Pedagogic application of regular expressions: A corpus-based online writing tool

```
/\bbetween\W+(?:\w+\W+){1,2}?to\b/gi;
```

John Blake
Center for Language Research, University of Aizu

Overview

Introduction:	context: universities
	situation and problem
	potential solutions
Literature:	errors
Corpus phase:	collection, annotation & analysis
Results:	error bank
Tool creation:	scripts, interface
Discussion:	user feedback
	next step

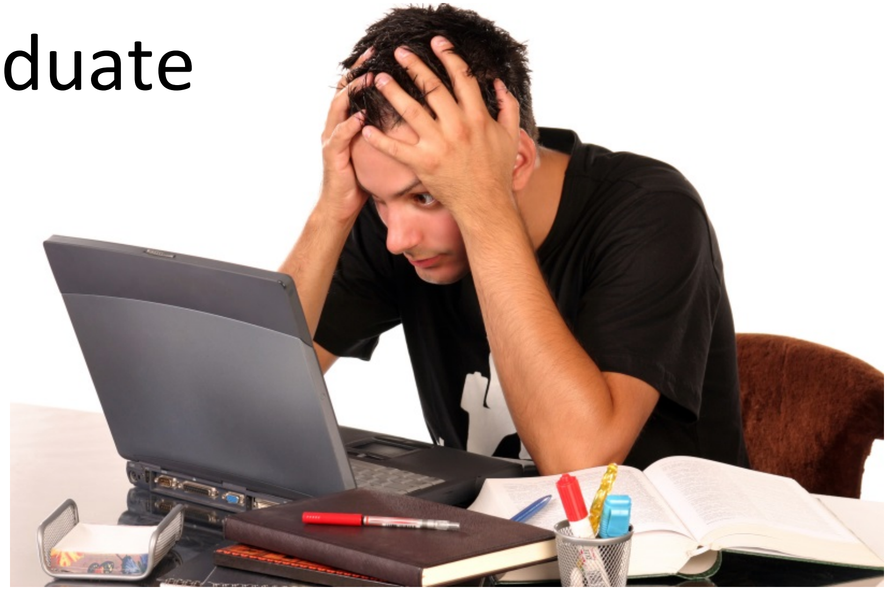
Context

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Problems:

Postgraduates (JAIST):
publish or don`t graduate

Undergraduates (Aizu):
draft thesis or don`t graduate



Teacher or student?

No time to respond

- Juggling research, lectures, admin, corpus symposiums, etc.

Piles of unmarked written work

1. Time eater
2. Predictable surface-level mistakes
3. Expectations of “correction”
4. Intended vs. perceived message

Reasons not to give feedback

- Feedback may not be read
- Feedback may be read but ignored
- Feedback may be misunderstood – especially pithy comments
- Lack of empirical evidence of benefits of providing feedback
- I cannot "mark" it yet because I haven`t finished my coffee

Pithy comments - examples

Original:

There are three main issue.

Feedback:

There are three main issue**s**.

There are three main **issue**.

There are three main issue. *****

There are **three** main **issue.** **Ag**

There are three main issue. **Ag**

Reasons to give feedback

Students expect teachers to:

- Identify the location of errors and/or
- Explain the errors and/or
- Correct the errors (if necessary)

All of which take lots of time.

What to give feedback on: Deep or superficial errors

- Respond to surface-level mistakes

  (easy for teacher and student)

- Respond to deeper mistakes

  (harder for teacher and student. Explanations are longer and rewrites are necessary)

Harness regex

Solution 1

Server-side script. Rule-based pattern matching

(2012 project)

Solution 2

Client-side script. Rule-based pattern matching

this presentation

Solution 3

Server-side script. Rule-based pattern matching and Probabilistic parsing future?

Blake, J. (2012, November 28-30). Corpus-based academic written error detector. *Conference proceedings of the 20th International Conference on Computers in Education.* Nanyang Technological University, Singapore.

Solution 2:
Rule-based pattern matching

1. Writer inputs text.
2. Text is searched.
3. Errors are identified.
4. Feedback is given for each error.
5. Students act on feedback.

Specific genre with high generic integrity (Bhatia, 1993)

- can target to user errors
- can r/o particular phraseologies unlike MS and Grammarly , e.g. There happened (to be a solution).

Solution 2:
Rule-based pattern matching

True/false statements

1. There is a man on your left. T / F

    If true, a man is on your left. Stop.

    If false, proceed to 2.

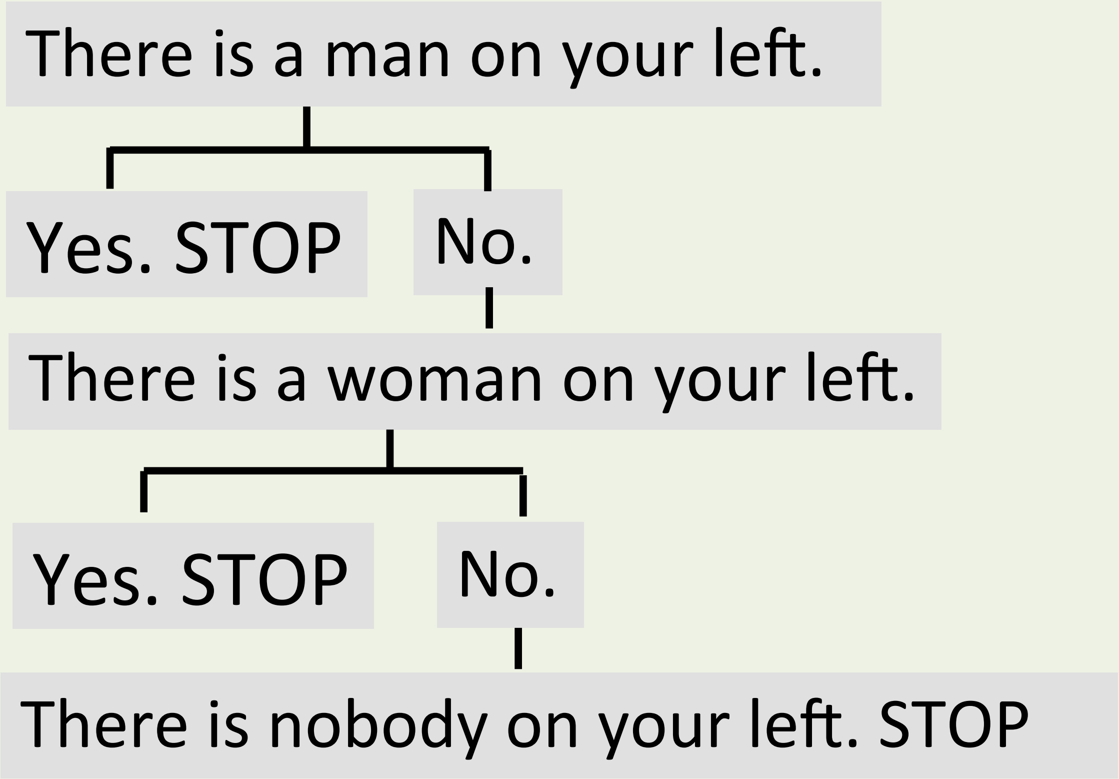
2. There is a woman on your left. T / F

    If true, there is a woman on your left. Stop.

    If false, there is nobody on your left. Stop.

Rule-based pattern matching

Decision-tree algorithm



There is a man on your left.

Yes. STOP

No.

There is a woman on your left.

Yes. STOP

No.

There is nobody on your left. STOP

Assumptions:

1. Only adults are present
2. There is no third gender

Rule-based pattern matching

Regular expressions (regexp|regex)

There is a man.	/\bman\b/;
There is a woman.	/\bwoman\b/;

The discrete words “man” and “woman” will be identified, generating a “true” result.

Rule-based pattern matching

Regular expressions (regexp|regex)

There is a man. /\bman\b/;

There is a woman. /\bwoman\b/;

The discrete words “man” and “woman” will be identified, generating a “true” result.

Regular expressions (Regex)

e.g. /\bmaybe\b/gi;

\ – escape (from normal characters)
i – case insensitive
b – boundary
g – greedy

1. I think that maybe he can understand. T/F
2. He may be able to understand T/F
3. Maybe, he can understand. T/F
4. Maybelline is a company name. T/F
5. Maybe, he said maybe. T/F

Types of language errors

Source

- Intralingual vs. interlingual errors (Selinker, 1972; Brown, 2000)
- Accidental slips, ingrained errors vs. attempts (Edge, 1990)
- Learner-induced vs. teacher-induced

Form and frequency

- Lexical, grammatical vs. discoursal
- Grammatical category (Orr & Yamazaki, 2004)

Effect

- Intrusive vs. non-intrusive errors
- Errors that lead to rejection vs. errors that don`t ← **My focus**

Brown , H. (2000). *Principles of Language Learning and Teaching.* New Jersey: Prentice Hall.

Edge, J. (1990). *Mistakes and correction.* Harlow: Longman.

Orr, T., & Yamazaki, A. K. (2004). Twenty problems frequently found in English research papers authored by Japanese researchers. *In Professional Communication Conference Proceedings International (pp. 23-35).*

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 209-231

Reasons for rejection/failure

Bordage, 2001; McKercher et al, 2007; Pierson, 2004; Thrower, 2012 and others report the main reasons as:

- Unoriginal
- Unimportant
- Flawed (method, analysis, etc.)
- Poor language ← **My focus**

Bordage G. (2001). Reasons reviewers reject and accept manuscripts: the strengths and weaknesses in medical education reports. *Acad Med, 76*(9), 889–896

McKercher B, Law R, Weber K, Song H, Hsu C (2007). Why referees reject manuscripts. *Journal of Hospitality & Tourism Research, 31*(4): 455-470

Pierson D.J. (2004). The top 10 reasons why manuscripts are not accepted for publication. *Respiratory Care, 49*(10): 1246-52.

Thrower, P. (2012). Eight reasons I rejected your article. Elsevier connect.

Method

Corpus collection

- 300 draft research articles (200 RA + 100 GT)
- Feedback given by tutors on articles was also collected

Corpus annotation

- To date around 4000 errors were annotated using Template analysis (King, 2004) with UAM Corpus Tool 3.0 (O'Donnell, 2015) *[stopped at 200 texts]*

Corpus analysis

- Frequency of occurrence
- Salience of errors *[code subject to funds]*

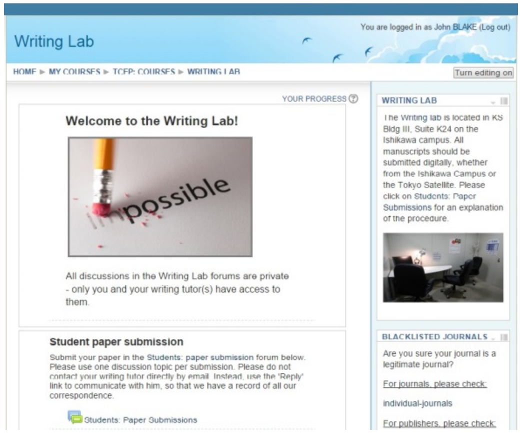
King, N. (2004). Using templates in the thematic analysis of text. In C.Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research (pp. 256–270).* London: Sage.

O'Donnell, M. (2015). UAM Corpus Tool (Version 3.0). Wagsoft Systems.

Corpus collection

Students submit article (& review comments).
Teachers provide feedback face-to-face.
Text converted to txt and added to corpus.

Online submission	Error identification	Writing consultations



You are logged in as John BLAKE (Log out)

Writing Lab

HOME ► MY COURSES ► TCFP: COURSES ► WRITING LAB

Welcome to the Writing Lab!

All discussions in the Writing Lab forums are private - only you and your writing tutor(s) have access to them.

Student paper submission

Submit your paper in the Students: paper submission forum below. Please use one discussion topic per submission. Please do not contact your writing tutor directly by email. Instead, use the 'Reply' link to communicate with him, so that we have a record of all our correspondence.

Students: Paper Submissions

WRITING LAB

The Writing lab is located in KS Bldg III, Suite K24 on the Ishikawa campus. All manuscripts should be submitted digitally, whether from the Ishikawa Campus or the Tokyo Satellite. Please click on Students: Paper Submissions for an explanation of the procedure.

BLACKLISTED JOURNALS

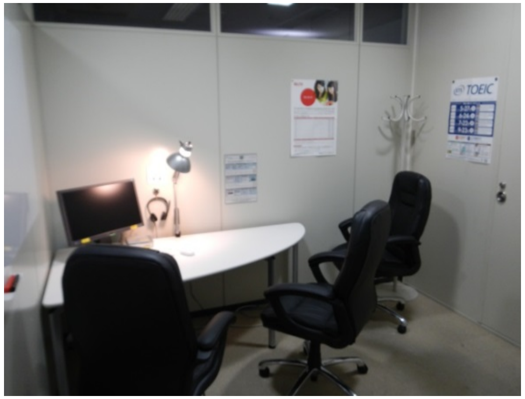
Are you sure your journal is a legitimate journal?

For journals, please check:

individual-journals

For publishers, please check:

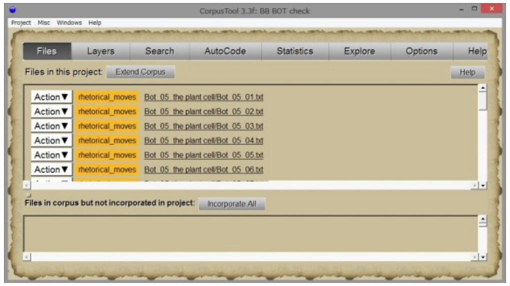
e.g. Highlight and number,
Insert comment,
Track changes,
Handwritten notes



Corpus annotation

Students submit article (& review comments).
Teachers provide feedback face-to-face.
Text is added to corpus.

UAM Corpus Tool



Files in this project:

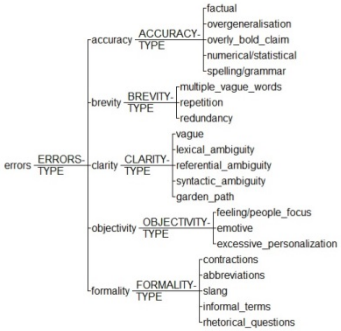
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_01.txt
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_02.txt
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_03.txt
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_04.txt
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_05.txt
Action▼ rhetorical_moves Bot_05_the plant cell/Bot_05_06.txt

Files in corpus but not incorporated in project:

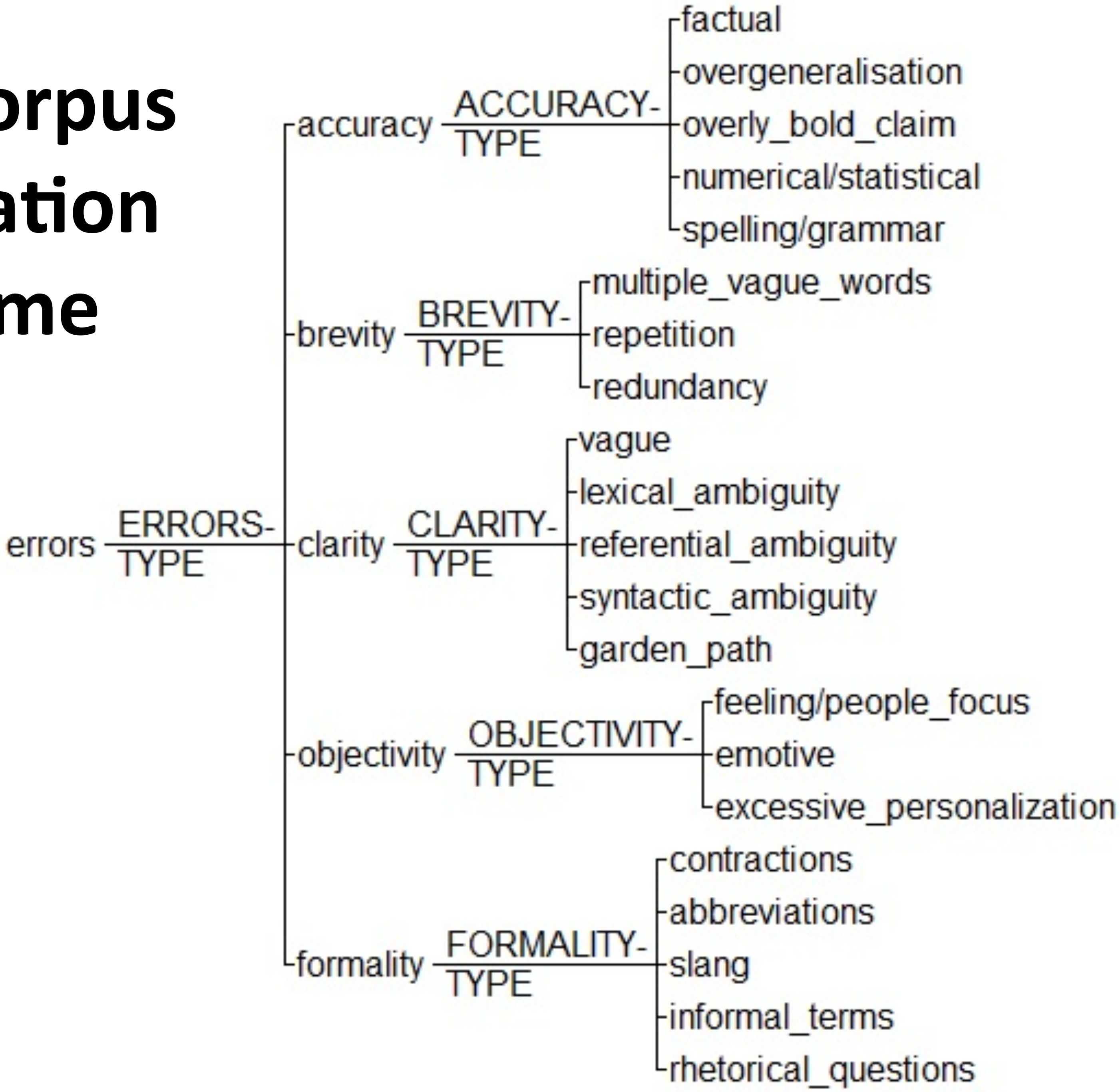
Evolution of ABCOF

Errors
→ ABC
→ ABCF
→ ABCOF

Refining sub-categories



Final corpus annotation scheme



- errors — ERRORS-TYPE
  - accuracy — ACCURACY-TYPE
    - factual
    - overgeneralisation
    - overly_bold_claim
    - numerical/statistical
    - spelling/grammar
  - brevity — BREVITY-TYPE
    - multiple_vague_words
    - repetition
    - redundancy
  - clarity — CLARITY-TYPE
    - vague
    - lexical_ambiguity
    - referential_ambiguity
    - syntactic_ambiguity
    - garden_path
  - objectivity — OBJECTIVITY-TYPE
    - feeling/people_focus
    - emotive
    - excessive_personalization
  - formality — FORMALITY-TYPE
    - contractions
    - abbreviations
    - slang
    - informal_terms
    - rhetorical_questions

Interim results

Corpus @ n=200 texts, single coder

- Five error types with 22 subgroups
- Grammatical accuracy errors were the most frequent (63%)
- Brevity (12%) and formality errors (11%) occurred
- Clarity (7%) and factual accuracy errors (3%) were less common but led to most confusion
- Objectivity errors were also infrequent (4%)

Code developed for common errors

Type	Typical problem areas
Accuracy*	mistakes in facts, meaning, grammar, usage and spelling
Brevity*	too many words to say something simple
Clarity*	vague or ambiguous terms
Objectivity	overly subjective terms
Formality	abbreviations, contractions and informal terms

* Initial coding used only three types.

Code developed for common errors

Type	Example errors
Accuracy	The population of Japan is **12,734,100** [1]
Brevity	…providing **the user** with various XXX and asking **him/her** to...
Clarity	Referring to Smith [10], Jones notes that **he**…
Objectivity	We are **confident** that XXX **will** become…
Formality	A **bunch** of IT engineers collaborated and launched…

Code developed for common errors

Type	Generic advice to avoid error
Accuracy	Check facts, spelling and grammar
Brevity	Remove redundancy
Clarity	Avoid ambiguity; be precise
Objectivity	Focus on things and ideas, not people and feelings
Formality	Avoid abbreviations, contractions and informal terms

Accuracy errors

1. The population of Japan is 12,734,100 [1].
2. There are two types of… First,.. Second, ..Third,..
3. All women …
4. XXX will play a key factor in the near future.
5. form XX to YY
6. $p <$ 0.5 cf. ($p < 0.05$) cf. ($p = 0.03$)

Accuracy errors

1. Factual errors related to world knowledge
2. Factual errors related to research topic
3. Overgeneralization errors
4. Overly bold claims
5. Spelling and grammar errors, esp. LaTeX users
6. Statistical errors

Brevity errors

1. The concept that was chosen as the primary focus of this research is XXX
2. ..providing the user with various XXX and asking him/her to XXX.
3. We analyze XXX regarding the XXX qualities, XXX qualities and XXX qualities.

Brevity errors

1. Using multiple vague words
2. Redundant words
3. Repeated words

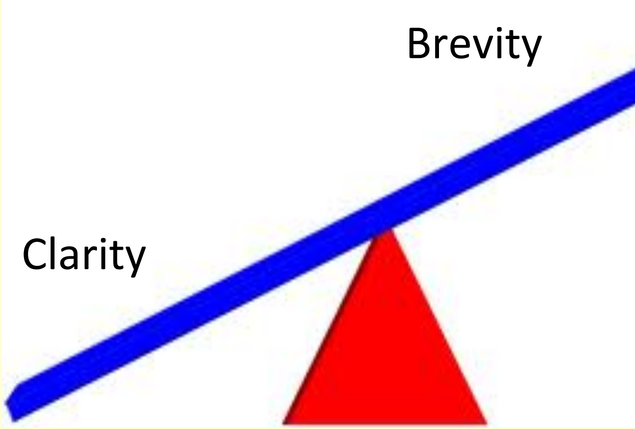
Clarity errors

1. XXX is something which is XXX from XXX of somewhere of, is something which XXX
2. It is really good for XXX.
3. Referring to Smith [10], Jones notes that he…
4. XXX found two AAA and one BBB, which CCC
5. The journal plans to publish this paper were just a rumour*

*not in corpus

Clarity errors

1. Vague expressions
2. Lexical ambiguity
3. Referential ambiguity
4. Syntactic ambiguity
5. Garden path sentences



Objectivity errors

1. We are confident that XXX will become XXXX
2. We are pleased to announce that XXX
3. …such as services to your XXX, to your XXX, and to XXX.

* ‘taming’ one’s subjectivity **(Peshkin,1988)**

Peshkin, A. (1988). In search of subjectivity. One`s own. *Educational Researcher, 17* (7), 17-21.

Objectivity errors

1. Focus on people & feelings, not things & ideas
2. Emotive wording
3. Excessive personalization, e.g. use of pronouns

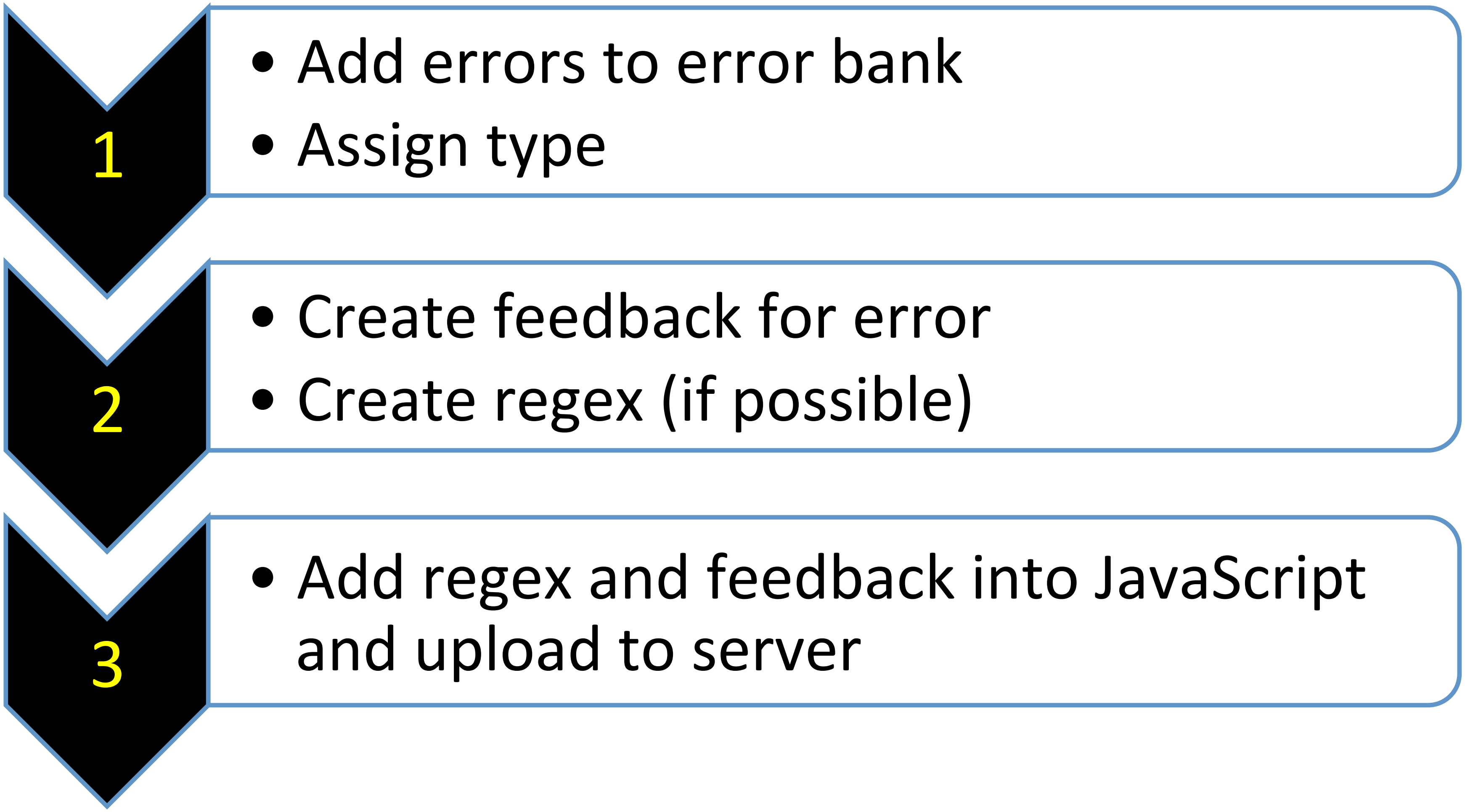
Formality errors

1. To be more precise, act doesn’t directly cause the effect (E).
2. This is the RQ of this paper.
3. A bunch of IT engineers collaborated and launched…
4. They launched the website right after the earthquake…
5. The key question to ask is: how can we…?

Formality errors

1. Contractions
2. Abbreviations
3. Slang
4. Informal terms
5. Rhetorical questions

Error to regex to detector



1

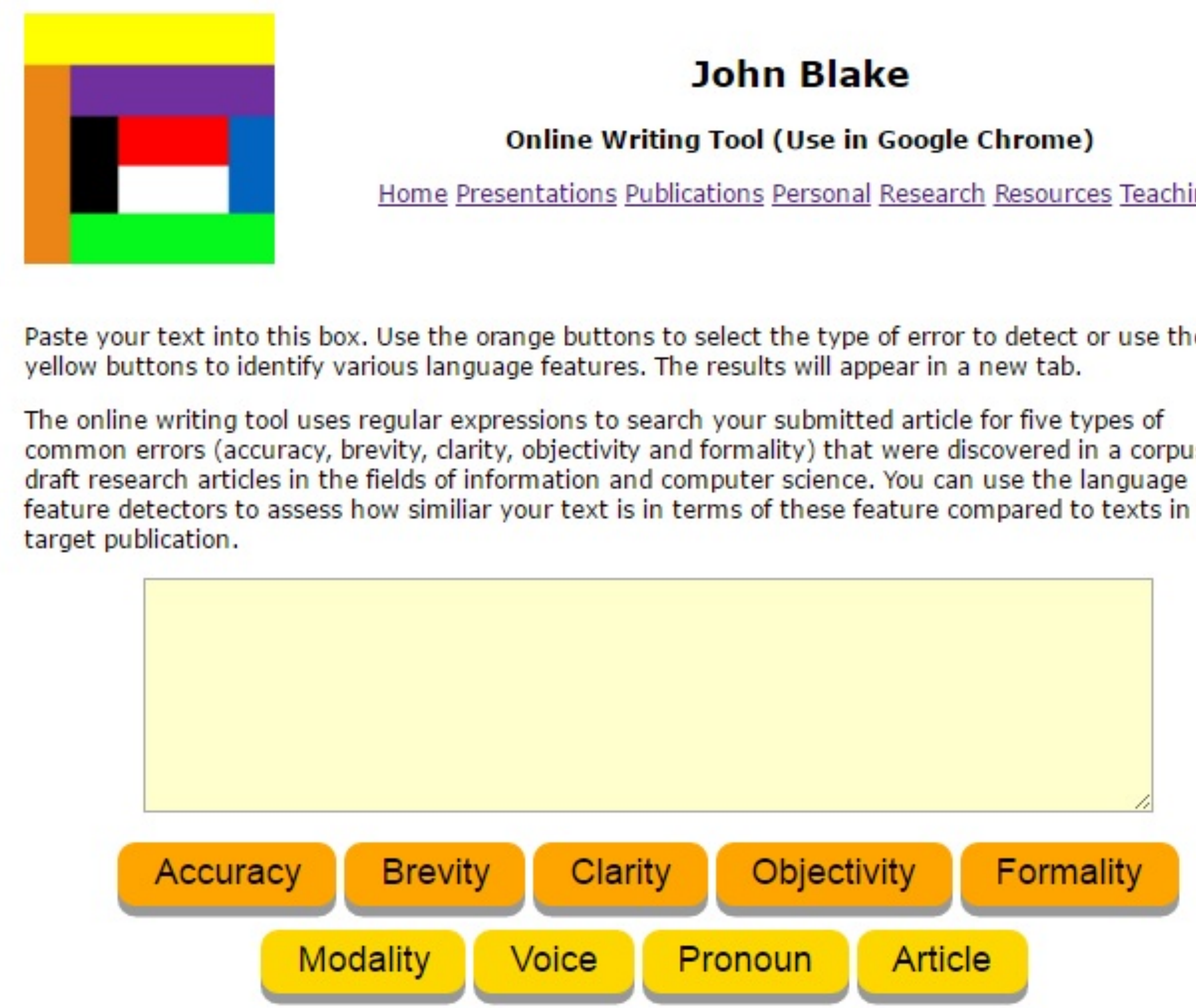
- Add errors to error bank
- Assign type

2

- Create feedback for error
- Create regex (if possible)

3

- Add regex and feedback into JavaScript and upload to server http://web-ext.u-aizu.ac.jp/~jblake/writingtool.html



John Blake

Online Writing Tool (Use in Google Chrome)

Home Presentations Publications Personal Research Resources Teaching

Paste your text into this box. Use the orange buttons to select the type of error to detect or use the yellow buttons to identify various language features. The results will appear in a new tab.

The online writing tool uses regular expressions to search your submitted article for five types of common errors (accuracy, brevity, clarity, objectivity and formality) that were discovered in a corpus of draft research articles in the fields of information and computer science. You can use the language feature detectors to assess how similiar your text is in terms of these feature compared to texts in your target publication.

Further development

Error tools	Transfer more regex from internal to external server
	Continue to add errors as corpus grows (until February)
Genre tools	SVOCA colour grammar (Patterns and language)
	Causality detector (Logic and language)

Initial time investment needed so cost-benefit assessment necessary. 10 students vs 200 students

Research

Text focus	1. Compare draft thesis to regex-checked thesis
	2. Compare regex feedback to actual alterations made in final version
Learning focus	3. Control vs Experimental group
	4. Qualitative study of users of tool

Conclusion

Benefit for students

- Legible and detailed feedback
- Easy to check with online dictionary
- Access 24/7 online

Benefit for teachers

- Reduces repetitive "correction"
- Time-saving so can focus on deeper learning (or research)

Initial time investment needed so cost-benefit assessment necessary. 10 students vs 200 students

Any questions, comments or suggestions?

jblake@u-aizu.ac.jp