Interdisciplinary Study

—the story of a computer scientist turned linguist or visa versa

Jason S. Chang

Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30

Waseda University, Tokyo

國立清華大學
National Tsing Hua University

Peeking into

the Future of Writing

—*Introducing WriteAway, an Interactive Writing Environment*

Jason S. Chang

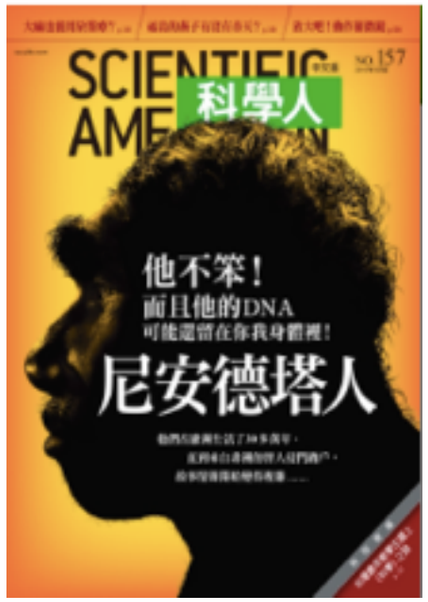
Department of Computer Science
National Tsing Hua University, TAIWAN

2015-0306 Friday 11:00-12:30

Waseda University, Tokyo

I wrote this month in Scientific American, Taiwan

I wrote this month in Scientific American, Taiwan



I wrote this month in Scientific American, Taiwan

- 雲端上的寫作教練 從泥版、竹簡、紙張到互動式寫作環境，寫作的面貌一變再變。

Always Online
網路不打烊

雲端上的寫作教練

從泥版、竹簡、紙張到互動式寫作環境，寫作的面貌一變再變。

電腦已能「閱讀」文章，而我們也瀏覽了電腦合成的時經新聞而不自知。

I wrote this month in Scientific American, Taiwan

- 雲端上的寫作教練 從泥版、竹簡、紙張到互動式寫作環境，寫作的面貌一變再變。
- Your Person Writing Trainer on the Cloud

  Many faces of writing: from Clay Tablet to Interactive Writing Environment (IWE)

How did Language emerge? evolve?

- Experts suggest that Language emerged (*The First Word* by *Christine Kenneally*)
  - Genes that control MOTION GRAMMAR (e.g., baby yawning)

- Motion of muscles --> motion of the vocal chord
- Language started with shapeless, volatile Speech
- Civilization needed something more concrete: Writing

- Language has been evolving much like plants and anim
- Writing changed more drastically than speech
  - Finger, stylus, quill, pen, ball pen
  - Clay, tablet, paper
- Today we use fingers
  - Keyboards, soft keyboards, touch screens
  - Writing appears in personal disk files, emails, blogs, Facebook (paperless)

New faces of Writing

- We have new ways of using Writing
  - Collaborative writing: Wikipedia
  - Community-based QA: Stackoverflow.com
  - Presenting yourself: Blogs
- Academic writing:
  - Marking Mate
  - Criterion
  - *WriteAway*

Writing programs in the 70s

- Writing in a more non-native language (e.g., Fortran, Pascal)
- Grammar is simple and well-defined
- When I learned how to program in college
- I wrote program using a card punching machine
- Computer read the cards
- One day later, an operator handed out the results
  - Syntax errors 401
- I punch a few new cards. And start over again

One more thing

- When people started to use this *new language*
- People were told to write Structural Programs
  - Use strictly 3 structures: sequence, condition, repetition
- Using a good programming language
  - Not Fortran (I would put C, C++ here)
  - Algol, Algol 68, Pascal, Python
- Go to's are considered harmful (Dijkstra 1968)
- Academic analog of programming
  - Rhetoric structure
  - Linking phrases

Programing in the past is like Writing today

- Students write in a non-native language (e.g., English)
- Grammar is complex and ill-defined
- Students write their essays in a word processor
  - Limited interactions (spelling, S-V agreement)
- Students hand in their assignment to the teacher
- One week later,
  - The teacher hands back graded papers, with
    - Grade
    - Corrective feedback (there is debate about this)

Writing programs TODAY

- People write programs in Interactive Development Environment
- In IDE
  - You type a few words (tokens)
    - The system will show you suggestions, what lies ahead
    - In other words, IDE help you with prompts and autocomplete
  - You *test drive* your program and get instant corrective feedback
    - Syntax errors
    - Semantic errors

My main point

- I will show you
  - **Future of Writing = Present of Programming**
  - **Future = Interactive Writing Environment**
- Why?
  - History repeats itself
  - We have better theory of the language
    - Lexical grammar
    - Pattern Grammar
  - We have much more Data to derive Pattern Grammar
  - Statistical methods are mature

Better Theory of Language

- Old theory: Two sides of a coin: Vocabulary + Rules
  - Words have parts of speech (POS)
    - Noun: object; Verb: action
    - Preposition/order: relation and event
  - Rules are related to POS, and almost independent of words
- New theory
  - Lexical grammar (e.g., Pattern Grammar)
  - Rules are intimately linked to words

Pattern Grammar

- PG is a model for describing the syntactic environments of individual words
- Each word has a set of patterns describing word usage in typical contexts
- One sense per pattern (often patterns are different for different word senses)

Sources:

- http://en.wikipedia.org/wiki/Pattern_grammar
- Hunston and Francis (2000): A corpus-driven approach to the lexical grammar of English

Pattern Grammar

Skim (v.) includes the following patterns in the COBUILD dictionary

- **V n off/from n**: *Skim the fat off the soup.* (limited prep. allow)
- **V n**: *Skim the wall surface smooth and ready for painting*
- **V over/across**: *Water skiers skimmed across the bay.*
- **V through n**: *Skim through the report and check for spelling mistakes?*

Sources:

- http://en.wikipedia.org/wiki/Pattern_grammar
- Hunston and Francis (2000): A corpus-driven approach to the lexical grammar of English

Dictionaries Embrace Pattern Grammar

- **have difficulty with something**: *She's having difficulty with her schoolwork this year.*
- **have difficulty (in) doing something**: *Six months after the accident, he still has difficulty walking.*
- **great/considerable difficulty**: *We had considerable difficulty finding anywhere to park.*
- **do something with/without difficulty**: *Seb was speaking with great difficulty.*

More Data Make It Easier to Derive PG

- **BNC**, *British National Corpus* 100 million words
- **COCA**, *Corpus of Contemporary American English* 450 million words
- **CiteseerX**, *Scholarly Big Data* 460 million words (2.8G)

Table 3: Collection and Usage Statistics

Statistic	Value
#Documents	3.5 million
#Unique documents	2.5 million
#Citations	80 million
#Authors	3-6 million
#docs added monthly	300,000
#docs downloaded monthly	300,000-2.5 million
Individual Users	800,000
Hits per day	2-4 million

Source: Wu, Zhaohui, et al. "Towards building a scholarly big data platform: Challenges, lessons and opportunities." in *Digital Libraries 2014*.

Take a closer look at CiteSeerX (for *WriteAway*)

$ head -3 citeseerx | tail | cut -f1

Because of their deployment in critical applications , the dependability modeling and analysis of Multiple-Phased Systems is a task of primary relevance .

$ wc citeseerx

20,000,000 sentences; 460,000,000 words; 2.8 G bytes

$ head -3 citeseerx | tail -1 | cut -f3,4

IN IN PRP$ NN IN JJ NNS , DT NN NN CC NN IN JJ NNS VBZ DT NN IN JJ NN .
(Part of Speech, read: prep, prep, possessive pronoun, ...)

I-PP H-PP I-NP H-NP H-PP I-NP H-NP O I-NP I-NP H-NP O H-NP H-PP I-NP H-NP H-VP I-NP H-NP H-PP I-NP H-NP O
(Base phrase, read: PP-start, PP -start, NP-start, NP-end, PP, NP-start, NP-end, ...)
PP = Prepositional Phrase, NP = Noun Phrase

How Big is the Scholar Data

- Three scholarly big data sets
  - Google Scholar
  - Microsoft Academic Search
  - Citeseer (mainly computer science)

Table 1: The Estimated Number of Scholarly Documents on the Web in Different Fields.

Discipline	Size in MAS	Estimated Size	public
Agriculture Science	447,134	1,088,711	12%
Arts & Humanities	1,373,959	5,286,355	24%
Biology	4,135,959	8,019,640	25%
Chemistry	4,428,253	10,704,454	22%
Computer Science	3,555,837	6,912,148	50%
Economics & Business	1,019,038	2,733,855	42%
Engineering	3,683,363	7,947,425	12%
Environmental Sciences	461,653	975,211	29%
Geosciences	1,306,307	2,302,957	35%
Material Science	913,853	3,062,641	12%
Mathematics	1,207,412	2,634,321	27%
Medicine	12,056,840	24,652,433	26%
Physics	5,012,733	13,033,269	35%
Social Science	1,928,477	6,072,285	19%
Multidisciplinary	9,648,534	25,798,026	43%
Total Sum		121,223,731	36,703,036

Entire con

Increasing Effective Statistical Tools

- Parsing and shallow parsing (based on the old theory)
  - *Genia tagger* (Tsuruoka et al. 2005)
- Collocation Extraction
  - *XTract* (Smadja 1993)
- Extracting Good Dictionary Examples
  - *GDEX* (Kilgarriff et al. 2008)
- Extracting Pattern Grammar from big data
  - *WriteAway* (Chang et al. 2015)

Source: Tsuruoka, Yoshimasa, et al. "Developing a robust part-of-speech tagger for biomedical text." *Advances in informatics*. Springer Berlin Heidelberg, 2005. 382-392.

My main point

- I will show you
  - **Future of Writing = Present of Programming**
- Why?
  - History repeats itself
  - We have better theory of the language
    - Lexical grammar
    - Pattern Grammar
  - We have much more data to derive Pattern Grammar

Steve Jobs famously said

iPod + phone + Internet communicator = iPhone

I'll say

Typewriter + Dictionary + Concordance = WriteAway

WriteAway

We have difficulty

less patterns | more patterns | less examples | more examples

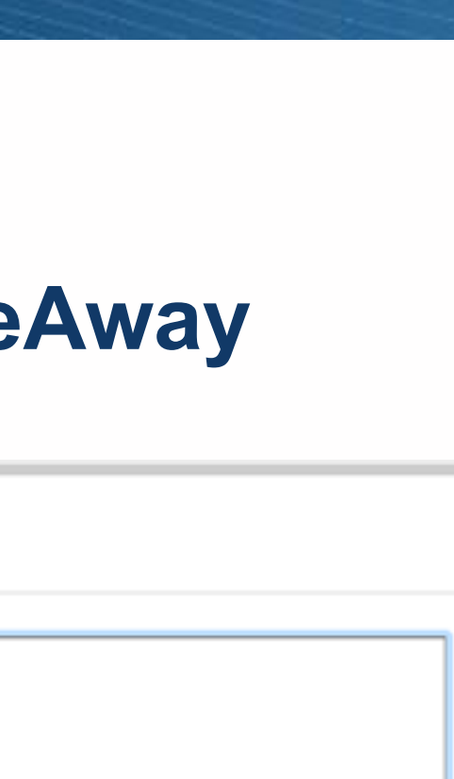
[N] difficulty with something 755

have difficulty with problems involving 25 5

difficulty with this approach/method is 82 13

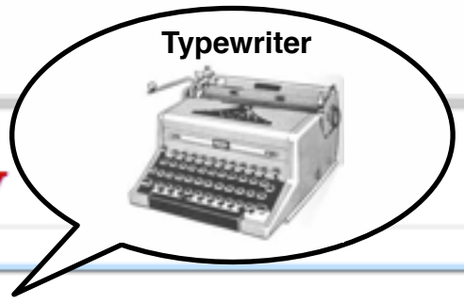
[N] difficulty doing something 718

have difficulty using it correctly 3 5



I'll say

Typewriter + Dictionary + Concordance = WriteAway



WriteAway

We have difficulty

less patterns | more patterns | less examples | more examples

[N] difficulty with something 755

have difficulty with problems involving 25 5

difficulty with this approach/method is 82 13

[N] difficulty doing something 718

have difficulty using it correctly 3 5

I'll say

Typewriter + Dictionary + Concordance = WriteAway

Typewriter

WriteAway

We have difficulty

Dictionary

less patterns more patterns ore examples

[N] difficulty with something 755

have difficulty with problems involving 25 5

difficulty with this approach/method is 82 13

[N] difficulty doing something 718

have difficulty using it correctly 3 5

I'll say

Typewriter + Dictionary + Concordance = WriteAway

Typewriter

WriteAway

We have difficulty

Dictionary

less patterns more patterns ore examples

[N] difficulty with something 755

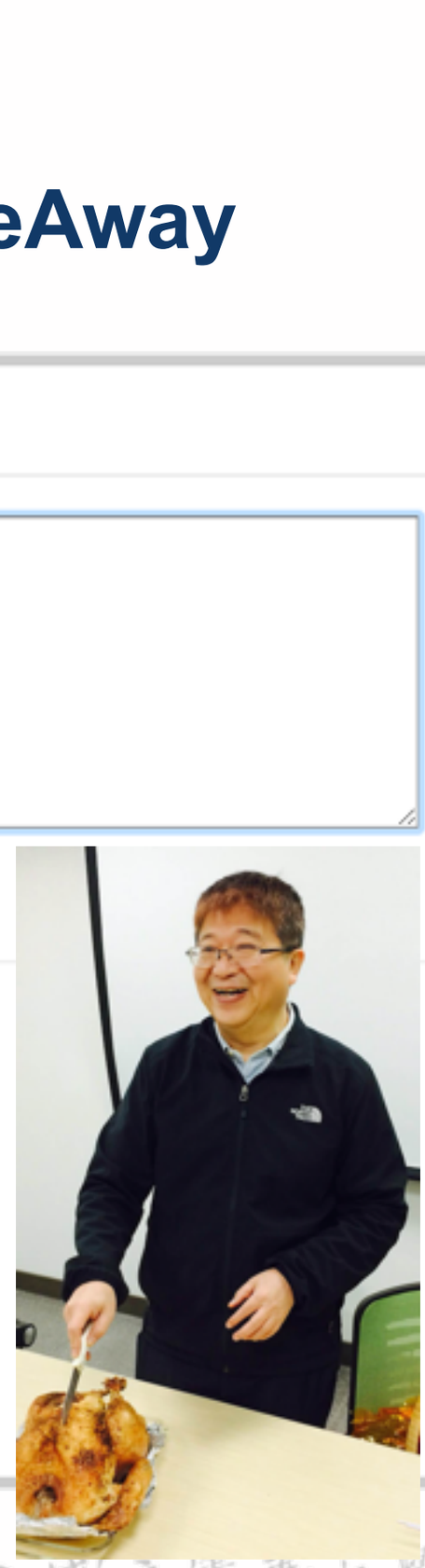
have difficulty with problems involving 25 5

difficulty with this approach/method is

Concordance

[N] difficulty doing something 718

have difficulty using it correctly 3 5



Concordance for Walden / Thoreau, Henry David

Use the features on this page to analyze and evaluate the text.

View full text
Words begining with:
Number of words:
Number of phrases:
Search:
Radius: Sort: None Left Right Match
Specialized searches: colors; adverbs; gerunds; "big names"; "great ideas"

1. ctualness of the civilized man? According to Liebig, man's b
2. ittle, while the civilized man hires his commonly because he
3. is tax, the poor civilized man secures an abode which is a p
4. commonly a poor civilized man, while the savage, who has th
5. tion between the civilized man and the savage; and, no doubt
6. ings. And if the civilized man's pursuits are no worthier th
7. contact with the civilized man. Yet I have no doubt that tha
8. a blessing. The civilized man is a more experienced and wis

WriteAway

This paper

less patterns | more patterns | less examples | more examples

[N] paper does 123527

paper presents/proposes a 35626 11894

paper describes/discusses the 24462 7026

[N] paper does something 84532

paper describes/presents a system 1960 72

paper presents an approach to 1219 113

WriteAway

This paper presents

less patterns | more patterns | less examples | more examples

[V] present something of something 52075

paper presents the results of a study conducted 630 7

paper presents the design and implementation of a mobile storage system called 259 6

[V] present something for doing something 27089

paper presents a parallel algorithm for solving the region growing problem based 235 7

we present a method for solving the following problem 142 11

WriteAway

This paper presents a method

less patterns | **more patterns** | **less examples** | more examples

[N] method for something 25928

is a multistart type stochastic ***method for bound constrained global optimization problems*** *1330 6*

by standard ***methods for one-dimensional systems*** *765 7*

[N] method be 24538

method is *applicable 17015 442*

[N] method of something 19839

method of analysis/proof *is 1325 115*

on the ***method of moments*** *433 20*

[N] method to do something 19659

of heuristic ***methods to solve hard computational search problems*** *528 8*

presents a single-pass , view-dependent ***method to solve the general rendering equation*** *160 8*

[N] method for doing something 18495

present a ***method for solving the following problem*** *960 11*

of numerical ***methods for solving ordinary differential equations*** *326 5*

Highlights of WriteAway

- Writing suggestions in the form of
  - grammar patterns
  - typical examples
  - Show less/more patterns/examples on demand
- Editing suggestions for the selected word
  - the same adjustable grammar patterns with examples
- Ranking patterns/examples according to context
  - move ‘difficulty doing something’ to the top when preceded by ‘have’
  - Attempt to resolve part of speech ambiguity (report n. vs report v.)
- Adjustable text box
  - You can type the whole paragraph

Advantages of Using WriteAway

- Countability
  - –Noticing the singular/plural forms in examples
- Verb tense/form
  - –Noticing the verb forms in examples
- Article
  - –Noticing the use or lack of articles
- Preposition
  - –Noticing n.+prep. or prep.+n. patterns
  - –E.g., evaluation on something
- Prep.+Verb Form
  - –Noticing grammar patterns
  - –E.g., method for doing something v.s. method to do something

What happens after this

- Improve WriteAhead
- Enter a competition to go to Silicon Valley Emersion Program
- Get more government funding (Taiwn MITI)
- Start a KickStarter campaign
- Collaborate Waseda University
- We are going after YOU

If you use Word, not getting enough help, why not use

WriteAway

to get a glimpse of the future of writing

If you use Word, not getting enough help, why not use

http://writeaway.nlpweb.org

to get a glimpse of the future of writing