LT RESEARCH FORUM
翻譯及語言學系講座

At City University of Hong Kong

Whether and How to Use Grammar in Grammatical Error Correction

Jason S. Chang
Department of Computer Science
National Tsing Hua University, Hsinchu, TAIWAN

2018 0320 16:00–17:30

Wearing Two Hats

- On my business card
  - I declare that I am a **computational linguist**
- Actually, I am wearing **two hats** in today's talk
  - Full-time **computer scientist**
  - Part-time **language teacher**

Source: www.teachingenglish.org.uk/article/error-correction

Whether to Use Grammar?

in Grammatical Error Correction

- Depending a lot on **the definition of** Grammatical Error Correction (GEC)
- Almost **absurd** to ask the question
- But, most state-of-the-art GEC systems use **little or no grammar**
- **Lack of grammar** makes it very difficult
  - to **characterize** the output beyond the obvious
  - to **explain** the output to the learner
  - to **improve** GEC systems

How to Use Grammar?

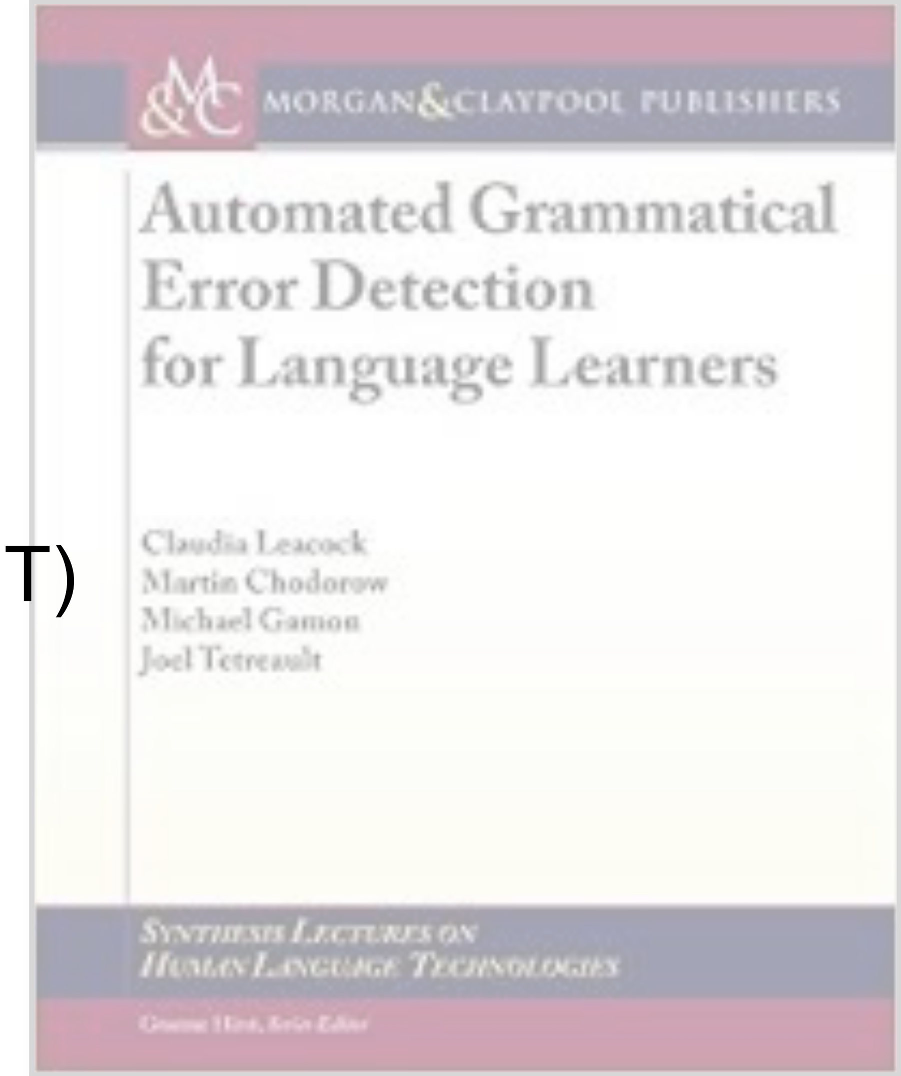
in Grammatical Error Correction

- **To annotate** the learner corpus (training data)
  - Currently, edit+pos are used (the Cambridge Scheme)
    - UT (unnecessary preposition)
    - e.g., We discussed [-about//UT] the issue
    - Different approach taken in Chinese Learner Corpus
- **To generate** artificial errors
  - Original: We discussed the issue
  - AGE: We discussed about the issue
- **To explain** GEC system output to a learner
  - Output: We discussed ~~about~~ the issue

    *You should delete **about** between **discussed** and **the***
  - GEC systems should do better than that

Defining GEC (as done *in Hard Science*)

- Definition in Leadcock et al. (2010)
  - Input: a sentence written by a learner
  - Output: a corrected sentence with errors marked
- Hard measure of success
  - Focusing on errors: Recall, Precision, $F_{.5}$
  - Focusing on sentences: GLEU
- Limitations
  - Assuming the results are used as is
  - Not discourse/history information (as in MT)
  - No teach-student / group dynamics



MORGAN & CLAYPOOL PUBLISHERS

Automated Grammatical Error Detection for Language Learners

Claudia Leacock
Martin Chodorow
Michael Gamon
Joel Tetreault

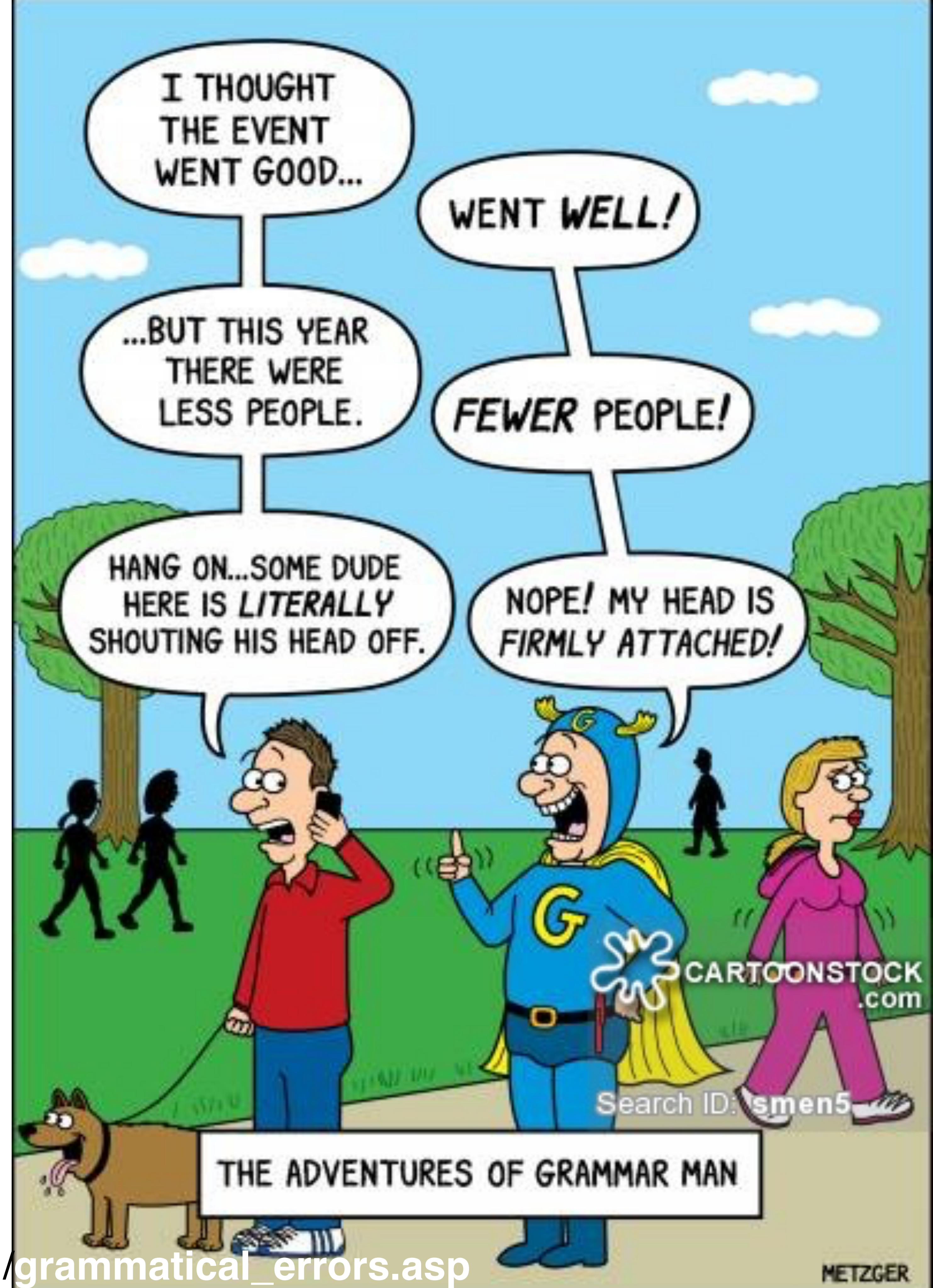
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, Series Editor

Grammatical Error Correction

in

Real Life


I THOUGHT THE EVENT WENT GOOD...
WENT WELL!
...BUT THIS YEAR THERE WERE LESS PEOPLE.
FEWER PEOPLE!
HANG ON...SOME DUDE HERE IS LITERALLY SHOUTING HIS HEAD OFF.
NOPE! MY HEAD IS FIRMLY ATTACHED!
G
CARTOONSTOCK .com
Search ID: smen5
THE ADVENTURES OF GRAMMAR MAN
METZGER

Source: www.cartoonstock.com/directory/g/grammatical_errors.asp

GEC seen *as a Soft Science*

- Dilemmas for teachers

  (*in* ***conversation*** *or* ***writing*** *groups*)

  - *it is always tricky to know*
    - ***when***
    - ***if*** *to correct students and*
    - ***how*** *to go about it*
- Questions and answers
  - *Don't* ***over-correcting***
  - *Do ask the students how* ***they want to be corrected***
  - *Focus on* ***accuracy*** *or* ***fluency****?*
  - ***Self*** */* ***Peer correction***
  - *When:* ***correction slots*** (‘**group hunt**) or ***on-the-spot*** *correction*
  - *Types:* ***New mistakes*** *or* ***old ones****?*

Jo Budden

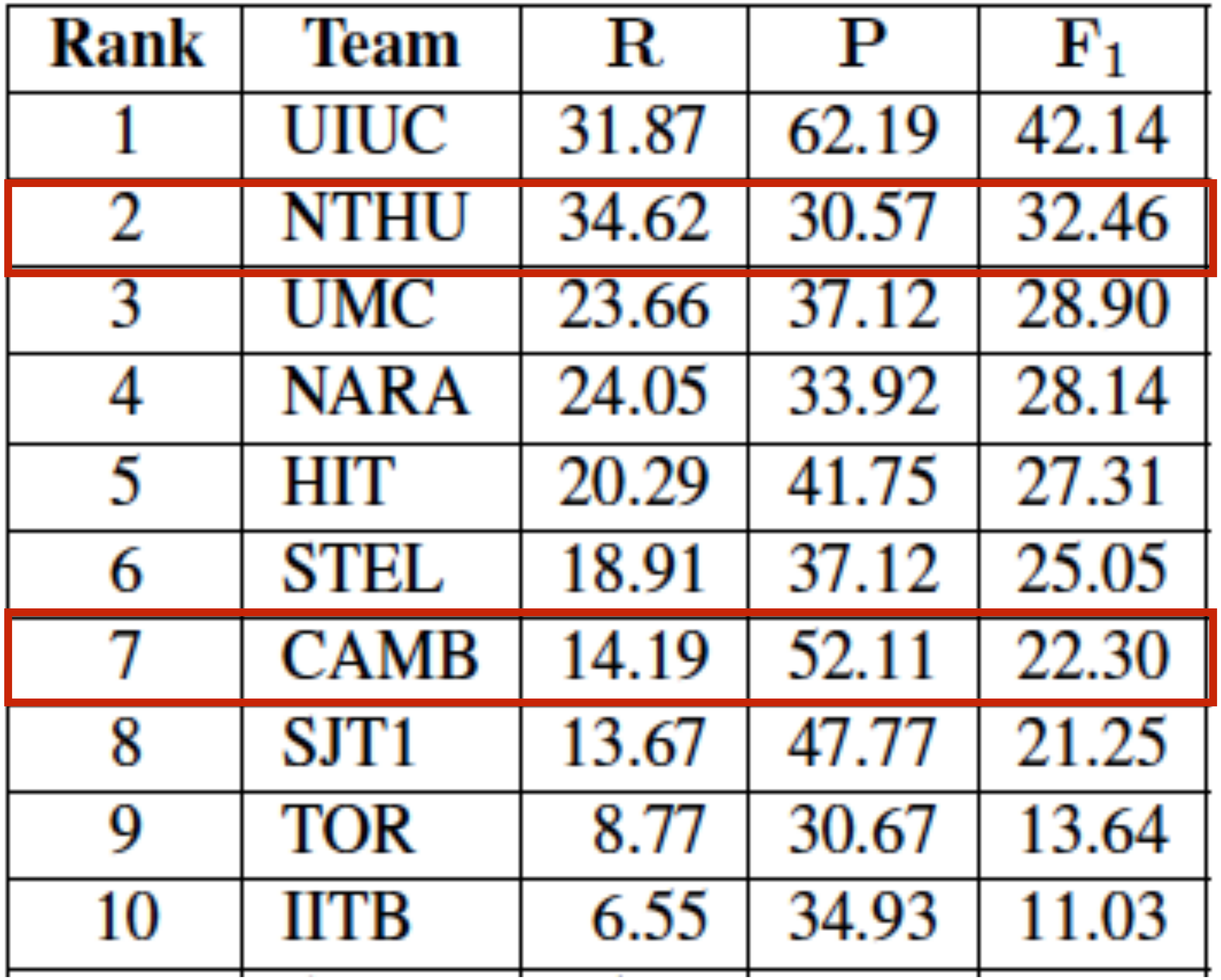
fequent logger on ***British Council BBC webpage***

Source: www.teachingenglish.org.uk/article/error-correction

History — Grammatical Error Correction

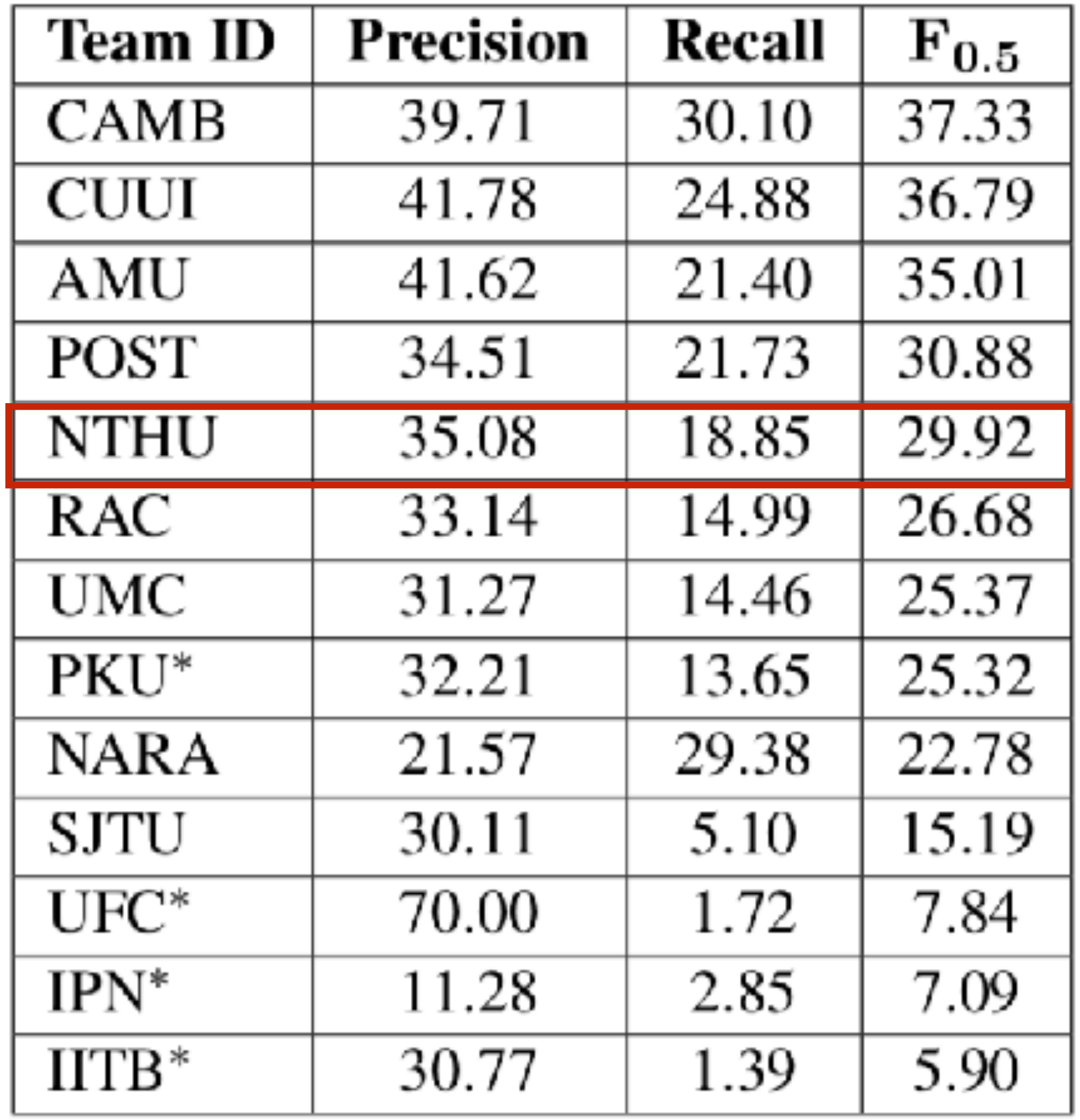
- 1945 ***Aspen*** software developed ***Grammatik***
- 1982 ***Heidorn and Jensen*** (IBM) applying parsing to develop ***Critique*** before moving to Microsoft
- 1992 ***Microsoft*** added a grammar checker to Word
- 2011 Help Our Own (HOO)
- 2013 CoNLL Shared Task: ***Big data stole the show***
- 2014 CoNLL Shared Task: ***SMT took over***
- 2016 Yuan at Cambridge adopted NMT: ***GEC went neural***
- 2017 Cambridge group ***augmented learner corpus with Artificial Error Generation (AEG)***
- 2018 Even with AEG, ***learner corpus*** could still be ***a source of problem***

2013 CoNLL Shared Task: *Big data stole the show*

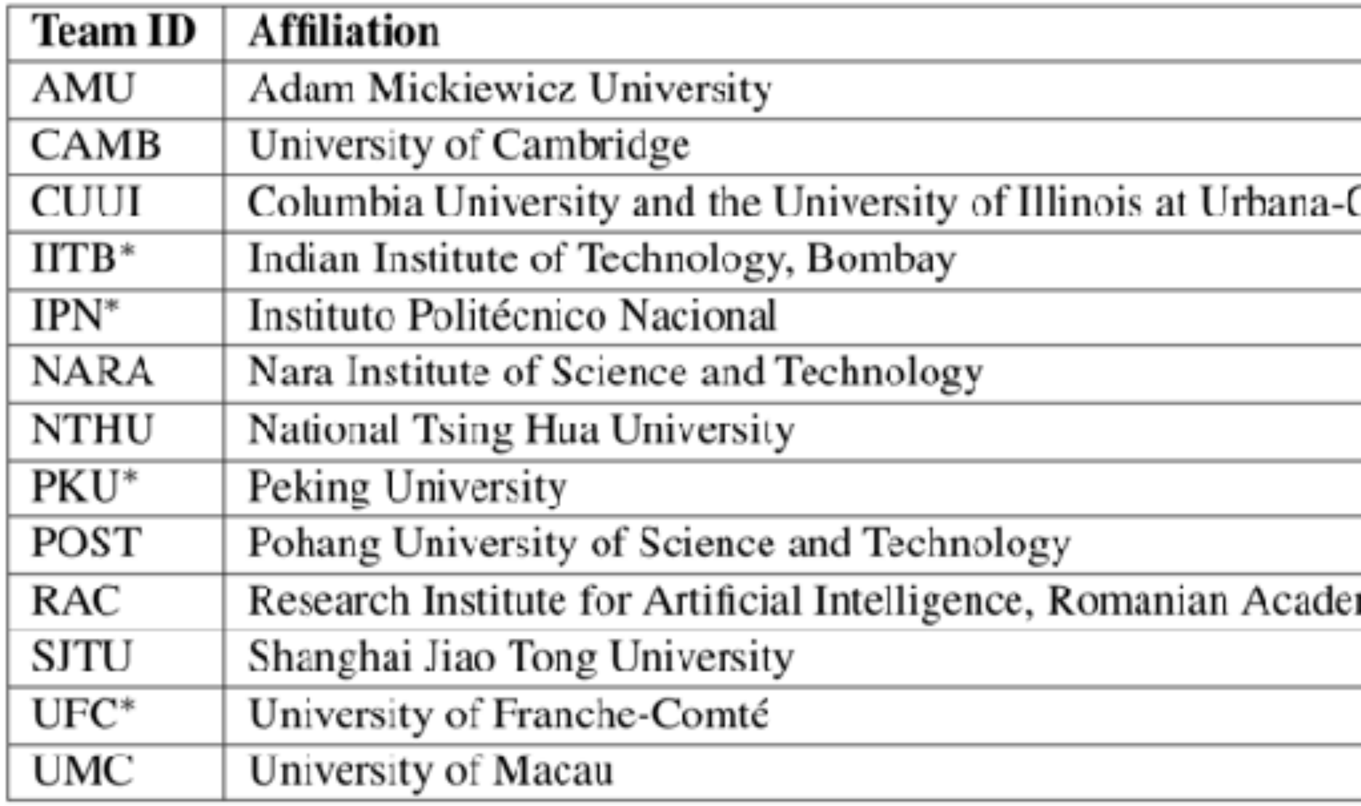


Rank	Team	R	P	F_1
1	UIUC	31.87	62.19	42.14
2	NTHU	34.62	30.57	32.46
3	UMC	23.66	37.12	28.90
4	NARA	24.05	33.92	28.14
5	HIT	20.29	41.75	27.31
6	STEL	18.91	37.12	25.05
7	CAMB	14.19	52.11	22.30
8	SJT1	13.67	47.77	21.25
9	TOR	8.77	30.67	13.64
10	IITB	6.55	34.93	11.03

2014 CoNLL Shared Task: *SMT took over*



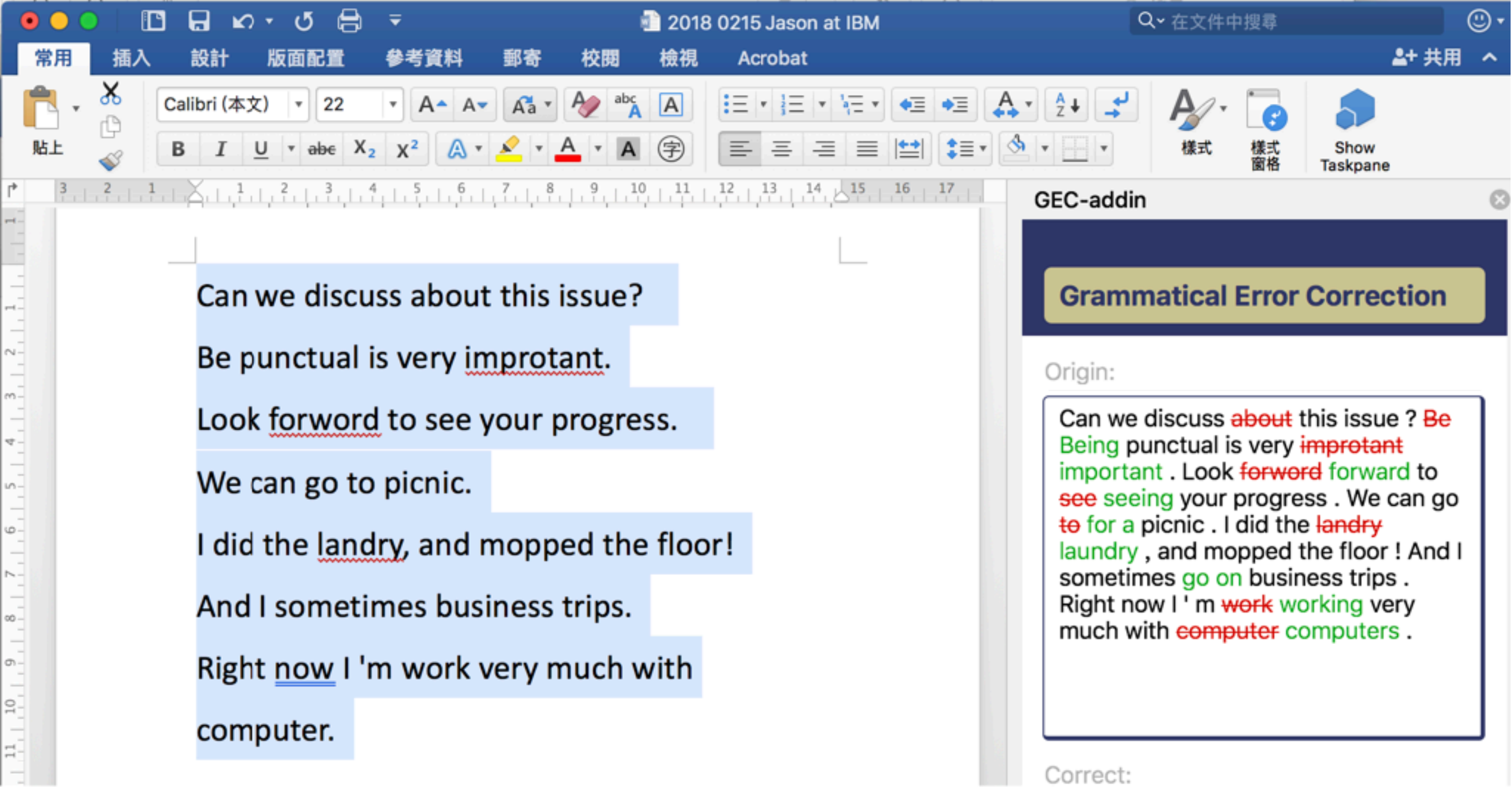
Team ID	Precision	Recall	$F_{0.5}$
CAMB	39.71	30.10	37.33
CUUI	41.78	24.88	36.79
AMU	41.62	21.40	35.01
POST	34.51	21.73	30.88
NTHU	35.08	18.85	29.92
RAC	33.14	14.99	26.68
UMC	31.27	14.46	25.37
PKU*	32.21	13.65	25.32
NARA	21.57	29.38	22.78
SJTU	30.11	5.10	15.19
UFC*	70.00	1.72	7.84
IPN*	11.28	2.85	7.09
IITB*	30.77	1.39	5.90



Team ID	Affiliation
AMU	Adam Mickiewicz University
CAMB	University of Cambridge
CUUI	Columbia University and the University of Illinois at Urbana-C
IITB*	Indian Institute of Technology, Bombay
IPN*	Instituto Politécnico Nacional
NARA	Nara Institute of Science and Technology
NTHU	National Tsing Hua University
PKU*	Peking University
POST	Pohang University of Science and Technology
RAC	Research Institute for Artificial Intelligence, Romanian Acade
SJTU	Shanghai Jiao Tong University
UFC*	University of Franche-Comté
UMC	University of Macau

Ongoing NLP Research at NTHU

- ***Grammatical error correction*** for English learners
  - Use Open NMT framework
  - Word add-in
- ***Spelling check for Chinese*** text
  - Use United Daily 50,000,000-word edit log
- ***Linggle*** (linguistic search engine)
  - Google Web 1T (1 trilling words, 5 gram)
  - CNA + United Daily 700+970 millions words
- ***WriteAhead*** (interactive writing environment)
- ***LanguageNet*** (mimicking ImageNet, BabelNet)
  - Sense-comparable multilingual examples/collocations in WordNet
  - Training data for word sense disambiguation and MT



Can we discuss about this issue?

Be punctual is very improtant.

Look forword to see your progress.

We can go to picnic.

I did the landry, and mopped the floor!

And I sometimes business trips.

Right now I 'm work very much with computer.

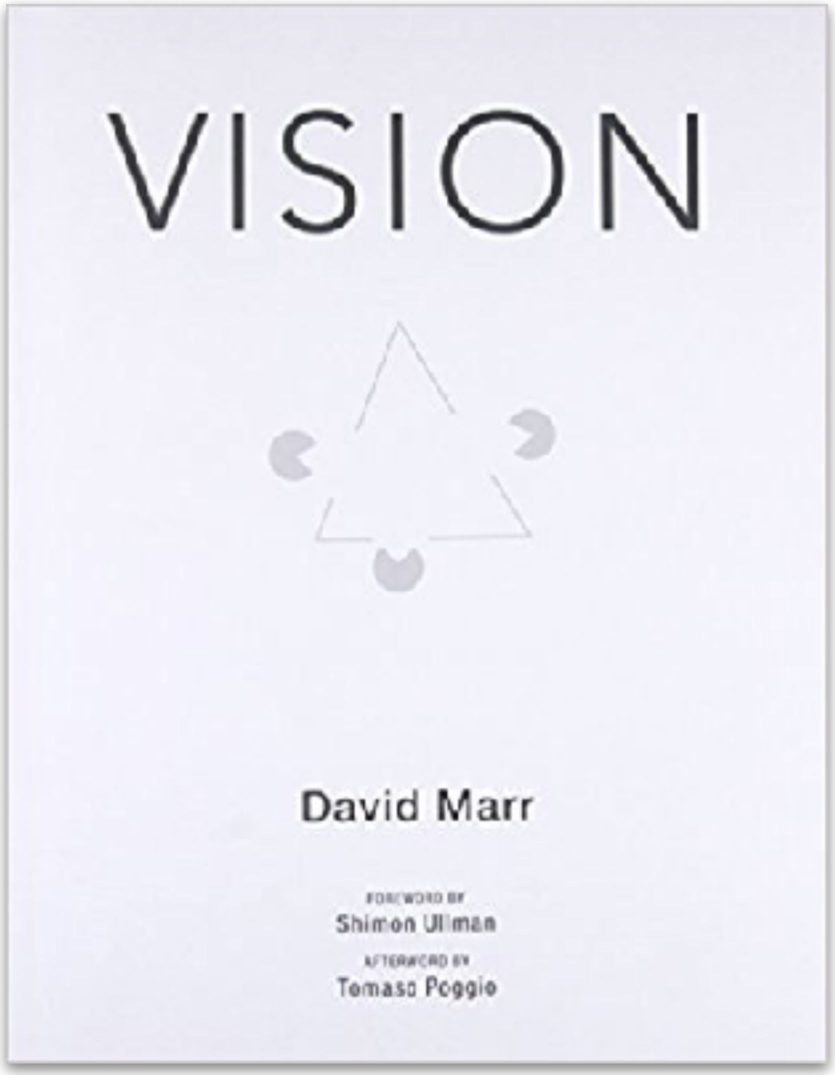
Grammatical Error Correction

Origin:

Can we discuss ~~about~~ this issue ? ~~Be~~ Being punctual is very ~~improtant~~ important . Look ~~forword~~ forward to ~~see~~ seeing your progress . We can go ~~to~~ for a picnic . I did the ~~landry~~ laundry , and mopped the floor ! And I sometimes go on business trips . Right now I ' m ~~work~~ working very much with ~~computer~~ computers .

Correct:

How is this done?



- Implementation
  - Topology: RNN, Bidirectional RNN
  - Recurrent unit: LSTM
  - Depth: 4 layers 500 hidden unit each
  - Optimizer: SGD
  - Training data: 2,200,000 sentences
  - Validation data: 243,191 sentences
  - Word vector size: 500, Batch size: 64, 13 epochs
- David Marr said it takes three levels to explain a complex system
  1. **Computational** (input vs. output)
  2. **Algorithmic** (procedure from IN to OUT)
  3. **Implementational** (coding of the procedure)

Input and Output in Grammatical error correction

- End-to-end and multiple grammatical errors types
  - Can we discuss ~~about~~ this issue? (preposition)
  - ~~Be~~ Being punctual is very ~~improtant~~ important. (verb form)
  - Look ~~forword~~ forward to ~~see~~ seeing your progress. (spell+verb form)
  - We can go ~~to~~ for a picnic. (prep.+det.)
  - I did the ~~landry~~ laundry, and mopped the floor! (spelling)
  - And I sometimes go on business ~~trip~~ trips. (missing+plural)
  - Right now I 'm ~~work~~ working very much with ~~computer~~ computers. (form+plural)
- Very little can be said in terms of Algorithm
  - words in represented as vectors
  - vectors were transformed/summed to regenerate the output

Limitations

- GEC is a problem desperately seeking a good dataset
  - EFCAMDAT is the best dataset, but still it is seriously faulted
    - <ability of -ing> errors are not consistent marked (increased 19 to 25 after human annotation)
    - same with <discuss about n> errors

le

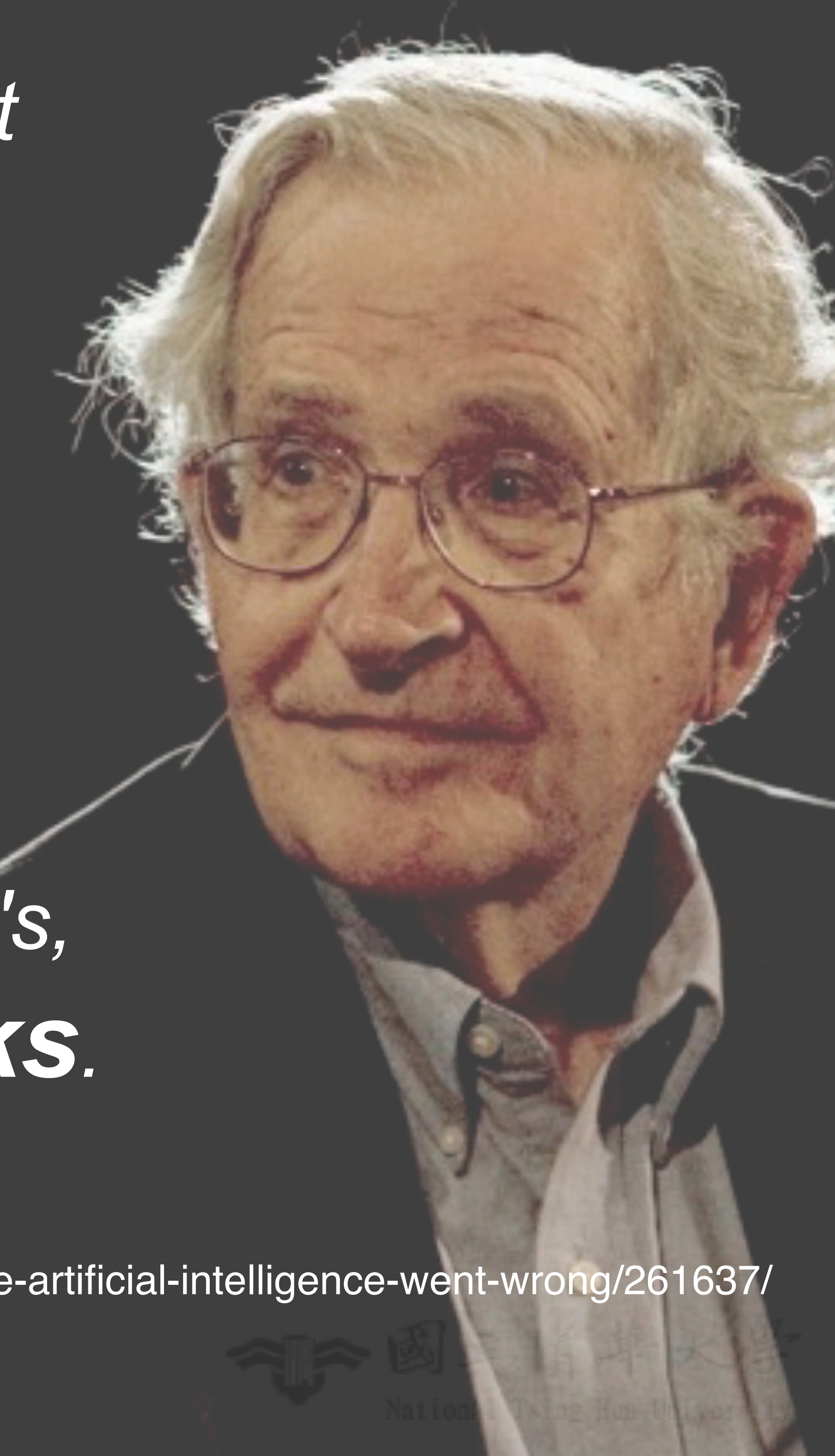
  - United Daily Edit Log is insufficient in coverage and generality (e.g., 今天 => 昨天, 造旨 => 造脂)
- Solution: Artificially Error Generation
  - How? Synchronous Pattern Grammar
    - ability: <N to-infinitive | N of -ing> (e.g., ability to think | ability of thinking)
      - original = What 's more , his **ability to speak** was perfect .
      - fake err = What 's more , his **ability of speaking** was perfect .
  - 中文 造詣 => 造旨 and 造詣 => 造藝

In the domain of biology, **would you consider** the work of **Mendel**, as a **successful** case? asked ***The Atlantics***

… Well, throwing out a lot of the data that didn't work.

[… But seeing the ratio that made sense, given the theory.]

*Yeah, he did the right thing. He **let the theory guide the data**. But that's, sure, that's **the way science works**. Same with chemistry.*



Source: www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/

The 701 Used Synchronous Grammar in Its Infancy

- The first Machine Translation System ***IBM Translator 701*** used “synchronous grammar” attached to words

. . . have taken normal ***words*** and ***attached*** to them ***tags or signs*** which give each word a precision it does not usually possess.

These signs actually ***denote rules of grammar and meaning***.

Although ***only six rules*** were used in today's demonstration … The six rules gover

[1] ***transposition of words*** where that is required in order to make sense,

[2] ***choice of meanings*** where a word has more than one interpretation,

[3] ***omission of words*** that are not required for a correct translation, and

[4] ***insertion of words*** that are required to make sense.

1957 Chomsky gave MT *Syntactic Structures*

one of the first serious attempts …

to construct … a comprehensive

theory of language …

in the same sense that a chemical,

biological theory is ordinarily

understood …

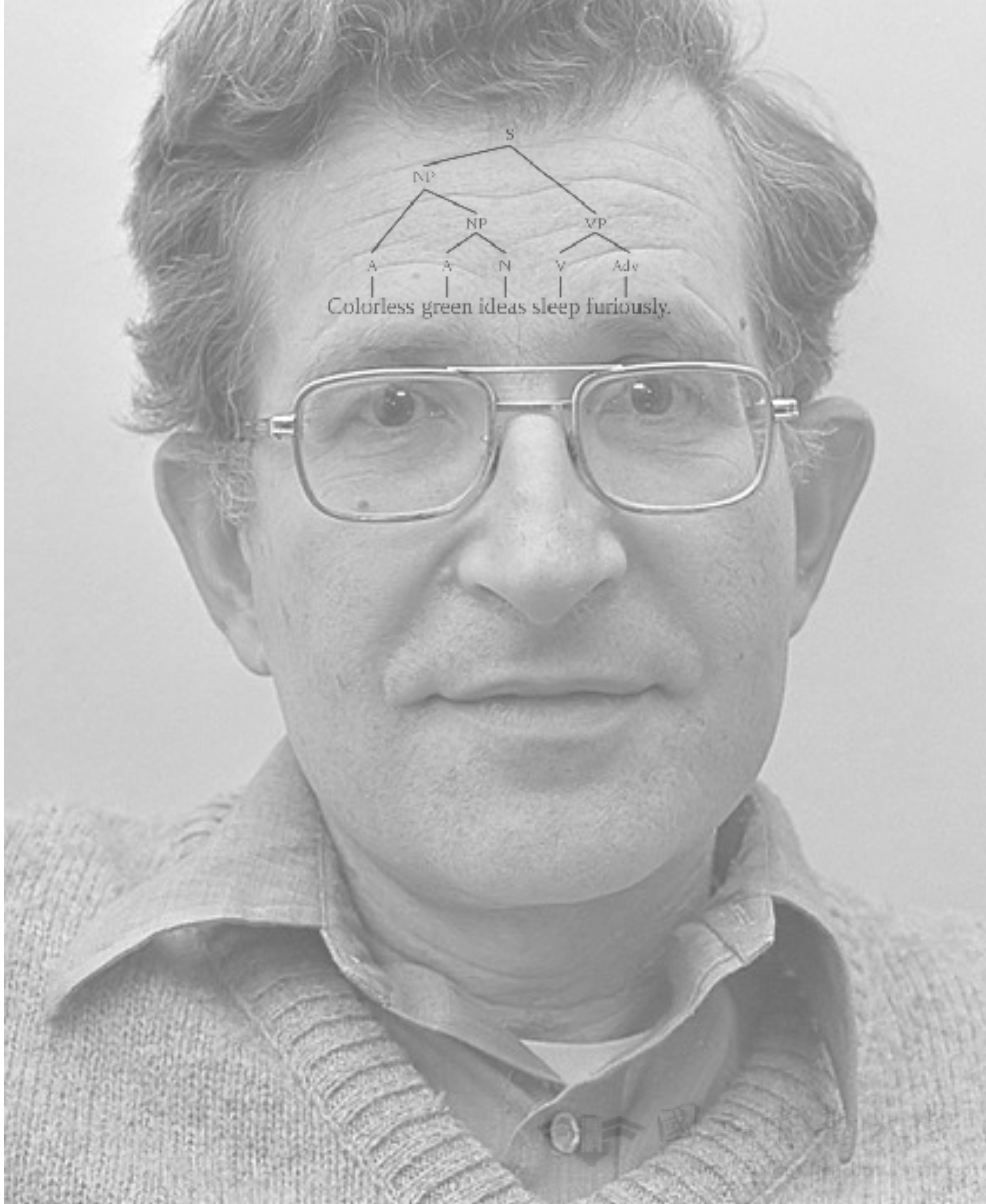
a rigorous explication of our

intuitions about our language in

terms of an overt axiom system,

the theorems …

— Robert B. Lees in *Language*

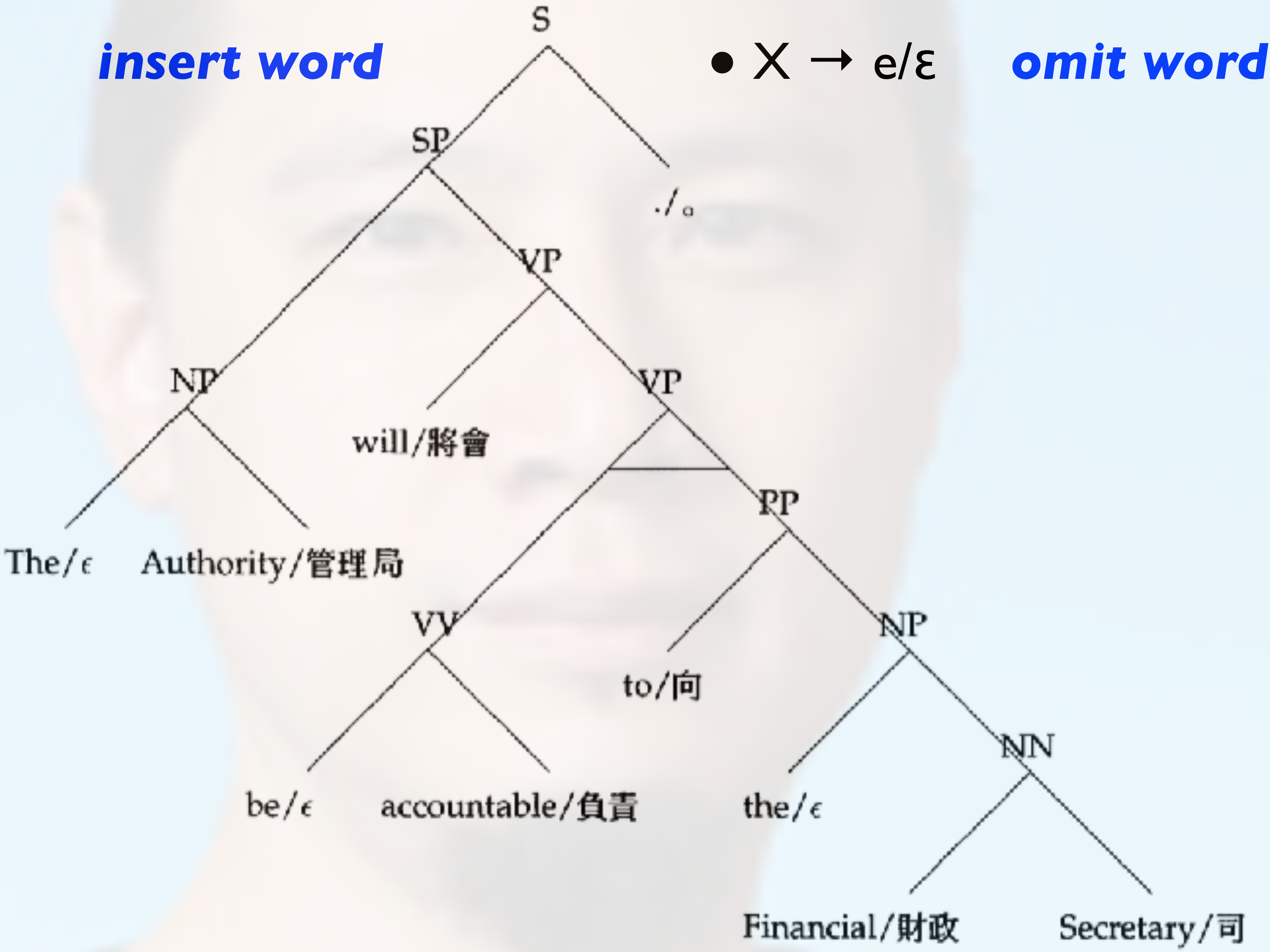


Source: en.wikipedia.org/wiki/Syntactic_Structures

40 Years later, Dekai Wu (1997) would rediscover . . .

Bracketing Transduction Grammar

- $X \rightarrow [X_1, X_2]$
- $X \rightarrow \langle X_1, X_2 \rangle$ ***transpose words***
- $X \rightarrow \varepsilon/f$ ***insert word***
- $X \rightarrow e/f$ ***word choice***
- $X \rightarrow e/\varepsilon$ ***omit word***



Solution: “He let the theory guide the data.”

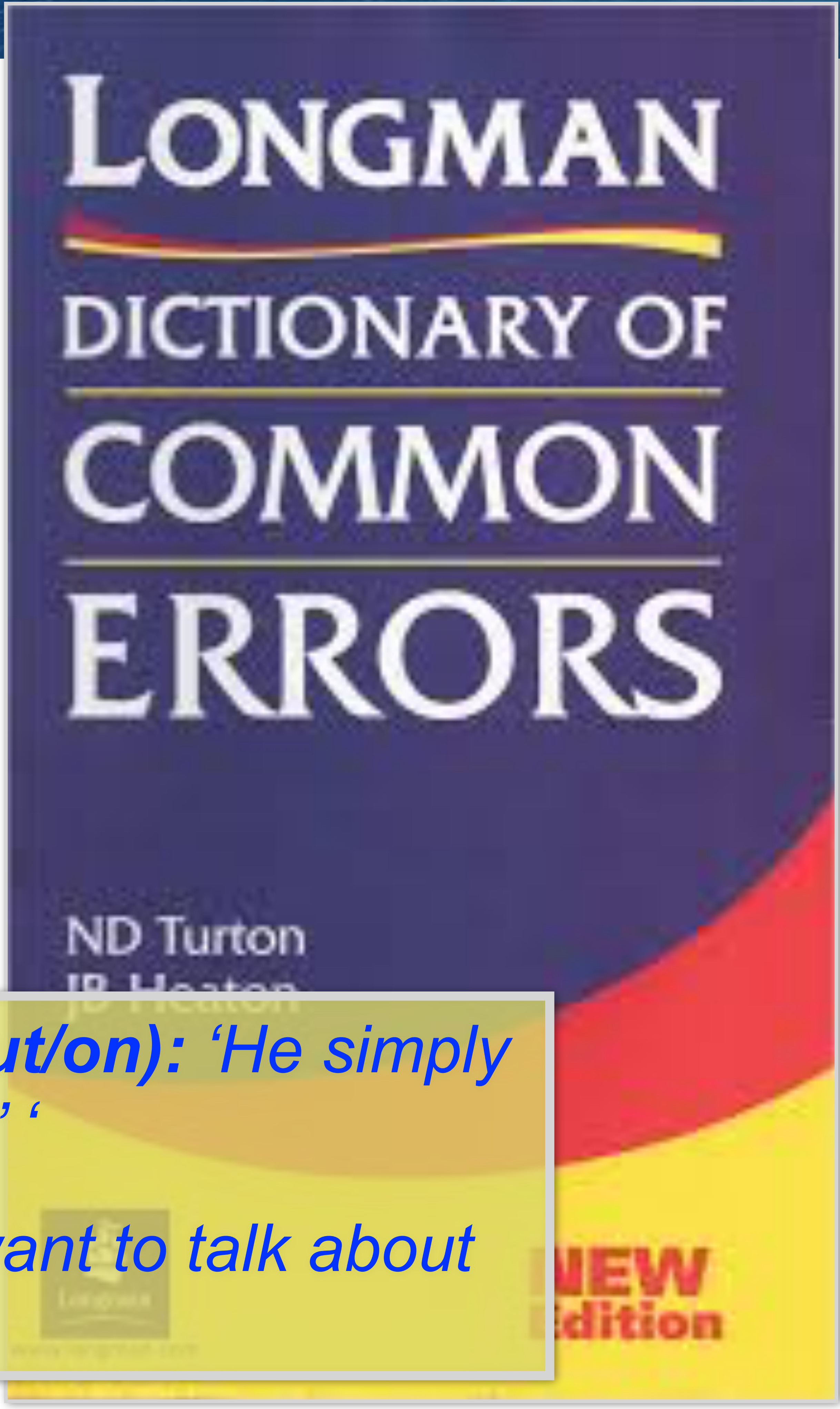
- In a way reminiscent of 701, consider “***discussed about the issues***”
- Attach to the word “***discussed***” the SPG rule [**V** ~~about~~ **n** | **V n**]
  (for ***omission of words***)
- Generate artificially errors using ***Synchronous Pattern Grammar***
  - **discuss**: <**V n** | **V about n**> (e.g [**discuss the issue** | **discuss about the issue**])
  - **ability**: <**N to-infinitive** | **N of -ing**> (e.g [**ability to think** | **ability of thinking**])
  - Apply SPG rules to “perfectly grammatical” sentences
    - original = What 's more , his **ability to speak** was perfect .
    - fake err = What 's more , his **ability of speaking** was perfect .
  - You have more data than you need to train a NMT-based GEC system
- This can be the basis for **XGEC**, eXplainable GEC, If you will.

XGEC: Explainable Grammatical Error Correction

- Recall ***Translator 701***
- Consider "***discussed about the issues***"
- Attach to the word "***discussed***" the SPG rule [**V** ~~about~~ **n** | **V n**] (for ***omission of words***)
- Key to giving the kind of explanations in ***LDOCE*** **(p 100)**

discuss sth (*WITHOUT* ***about/on):*** *'He simply refuses to discuss the matter.' '*

Compare ***talk about:*** *'They want to talk about what to do next.'*



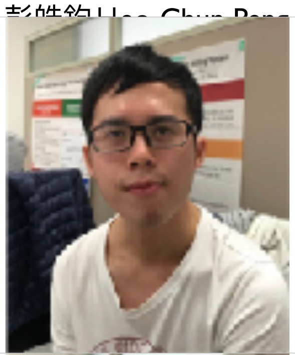
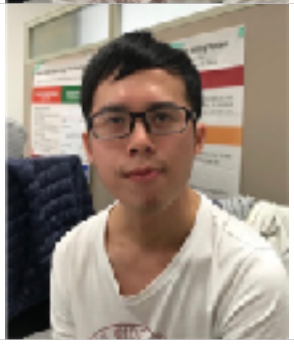
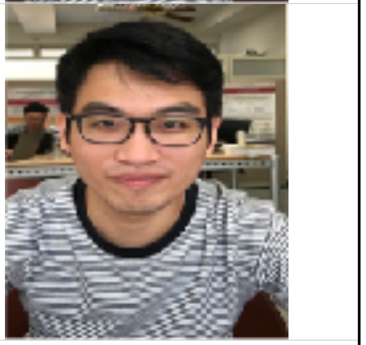
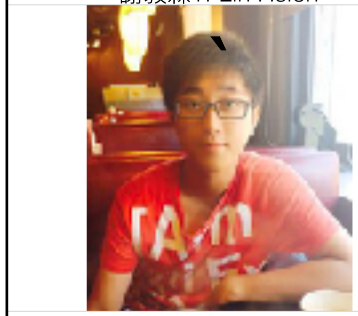
LONGMAN
DICTIONARY OF
COMMON
ERRORS

ND Turton
JB Heaton

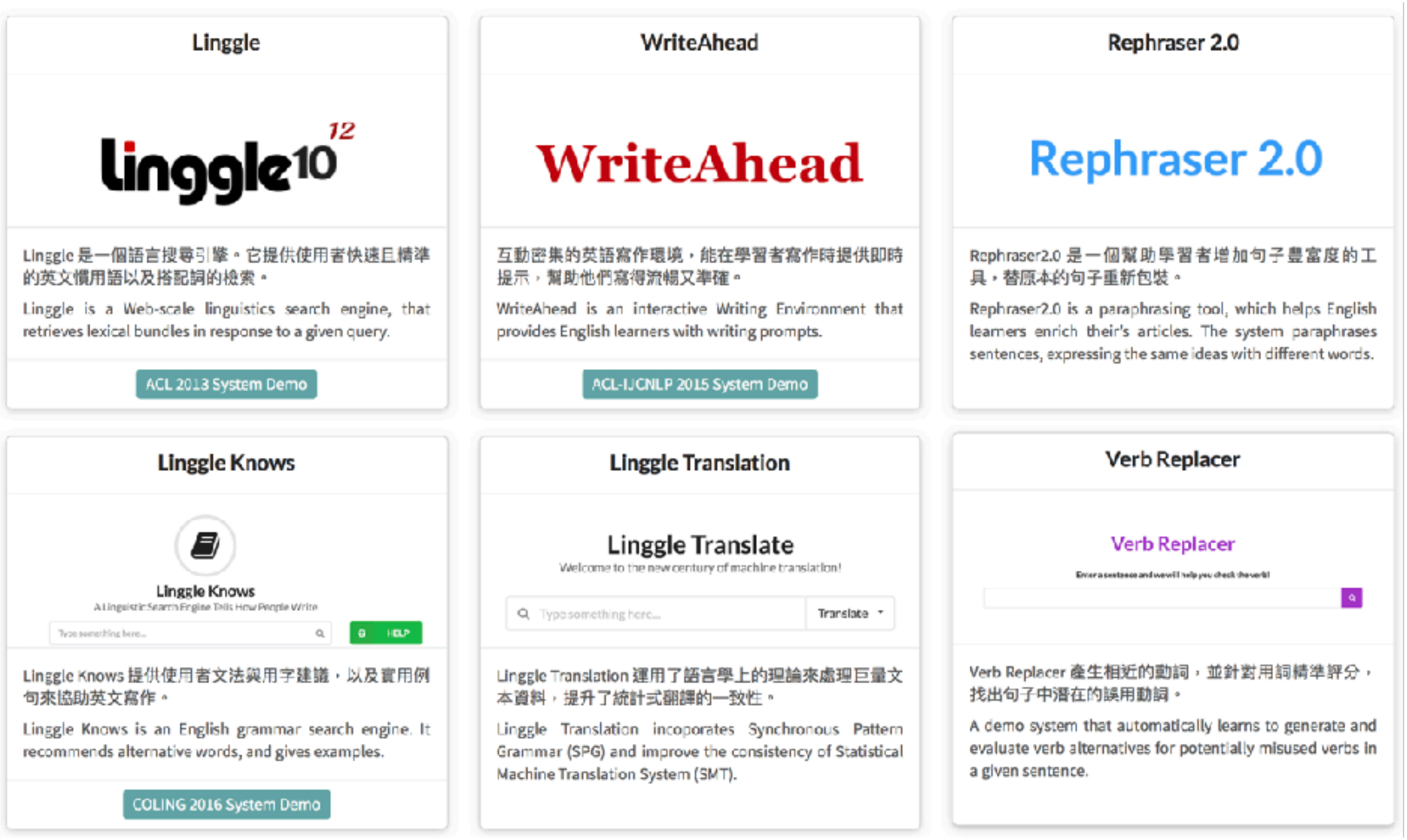
NEW Edition

Natural Language Group

National Tsing Hua University

張俊盛Jason S. Chang	楊馨瑜Ching-Yu Yang	羅曼Frances M. Lo	蓋嘉瑋Chia-Wei Kai	彭皓鈞Hao-Chun Peng	朱鳳華Huang-Hua Ju
					
					
陳志杰Jhih-Jie Chen	李巧雯Chiao-Wen Li	程尚謙Shang-Chien Cheng	羅右鈞Yu-Chun Lo	彭成全Chen-Quen Peng	林昌毅Chung-I Lin
謝毅霖Yi-Lin Hsieh	韓文彬Wen-Bin Han	蔡仲庭Chung-Ting Tsai	蔡名喬Ming-Chiao Tsai	何佳芳Chia-Fang Ho	Joanne Boisson
					
黃翰	陳品媛 Jessica	羅婕瑜 Je Yu	Sunny	Wanyu	Nica

We published our systems in ACL, NAACL, EMNLP, Coling, IJCNLP


linggle10 12
WriteAhead
Rephraser 2.0
Linggle Knows
A Linguistic Search Engine Tells How People Write
Type something here...
HELP
Linggle Translate
Welcome to the new century of machine translation!
Type something here...
Translate
Verb Replacer
Enter a sentence and we will help you check the verb!

Linggle

Linggle 是一個語言搜尋引擎。它提供使用者快速且精準的英文慣用語以及搭配詞的檢索。

Linggle is a Web-scale linguistics search engine, that retrieves lexical bundles in response to a given query.

ACL 2013 System Demo

WriteAhead

互動密集的英語寫作環境，能在學習者寫作時提供即時提示，幫助他們寫得流暢又準確。

WriteAhead is an interactive Writing Environment that provides English learners with writing prompts.

ACL-IJCNLP 2015 System Demo

Rephraser 2.0

Rephraser2.0 是一個幫助學習者增加句子豐富度的工具，替原本的句子重新包裝。

Rephraser2.0 is a paraphrasing tool, which helps English learners enrich their's articles. The system paraphrases sentences, expressing the same ideas with different words.

Linggle Knows

Linggle Knows 提供使用者文法與用字建議，以及實用例句來協助英文寫作。

Linggle Knows is an English grammar search engine. It recommends alternative words, and gives examples.

COLING 2016 System Demo

Linggle Translation

Linggle Translation 運用了語言學上的理論來處理巨量文本資料，提升了統計式翻譯的一致性。

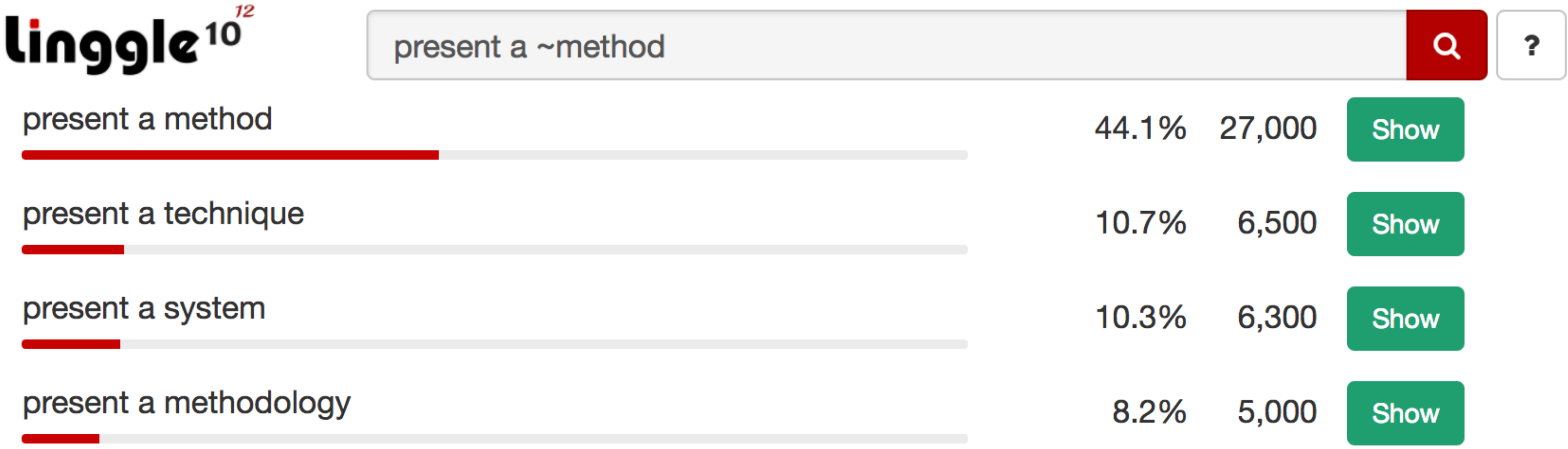
Linggle Translation incoporates Synchronous Pattern Grammar (SPG) and improve the consistency of Statistical Machine Translation System (SMT).

Verb Replacer

Verb Replacer 產生相近的動詞，並針對用詞精準評分，找出句子中潛在的誤用動詞。

A demo system that automatically learns to generate and evaluate verb alternatives for potentially misused verbs in a given sentence.

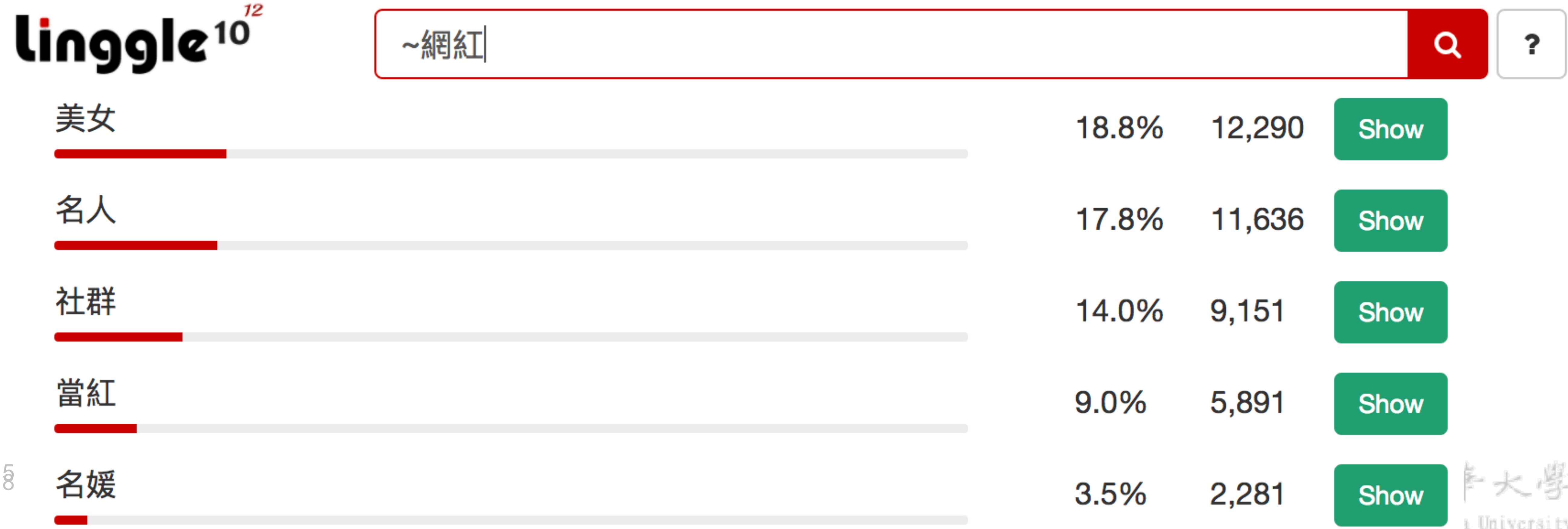
Linggle English and Chinese Versions



present a ~method

present a method	44.1%	27,000	Show
present a technique	10.7%	6,500	Show
present a system	10.3%	6,300	Show
present a methodology	8.2%	5,000	Show

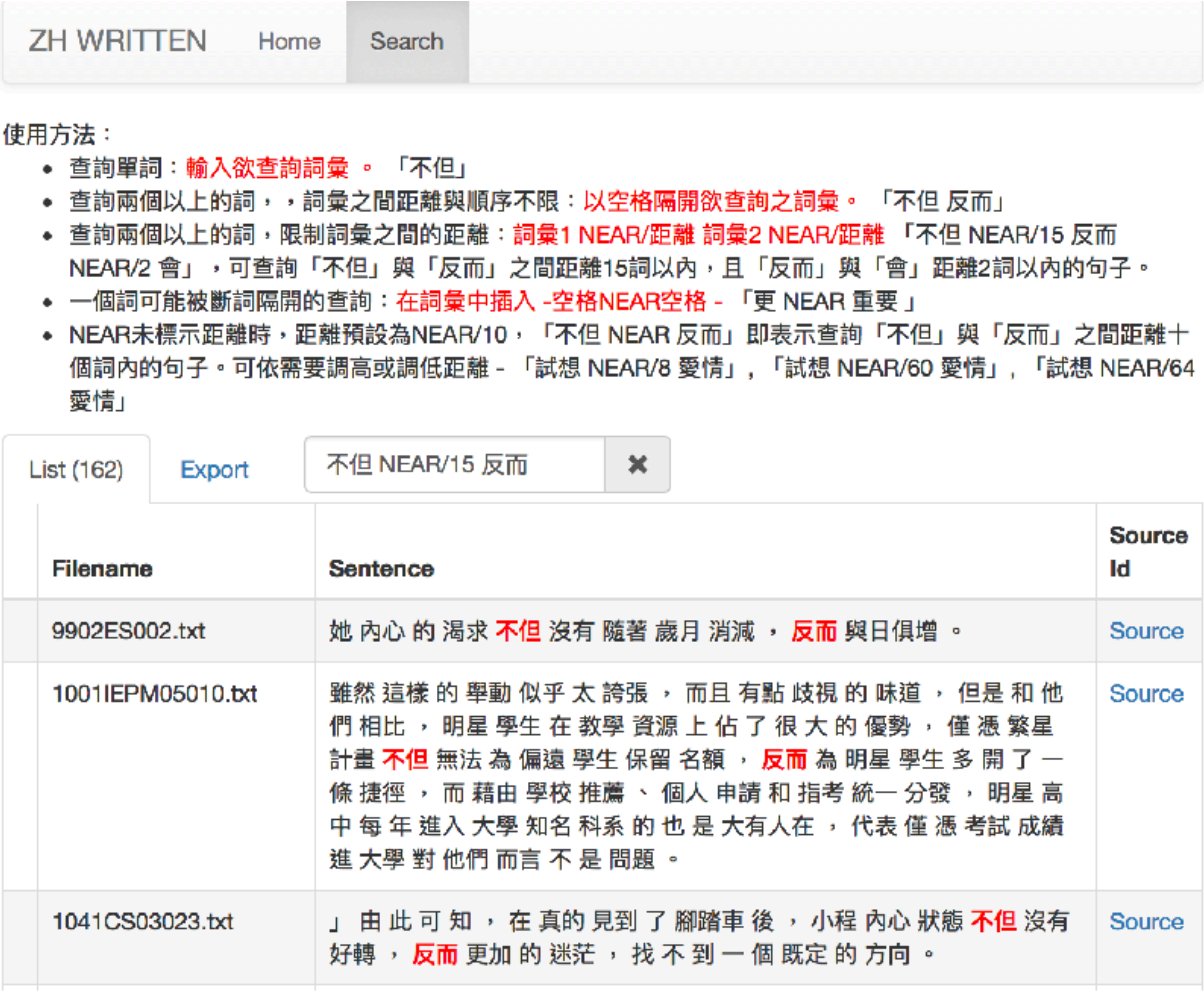
le



~網紅

美女	18.8%	12,290	Show
名人	17.8%	11,636	Show
社群	14.0%	9,151	Show
當紅	9.0%	5,891	Show
名媛	3.5%	2,281	Show

Concordance of NTHU Chinese Learner Corpus



使用方法：

- 查詢單詞：輸入欲查詢詞彙 。 「不但」
- 查詢兩個以上的詞，，詞彙之間距離與順序不限：以空格隔開欲查詢之詞彙。 「不但 反而」
- 查詢兩個以上的詞，限制詞彙之間的距離：詞彙1 NEAR/距離 詞彙2 NEAR/距離 「不但 NEAR/15 反而 NEAR/2 會」，可查詢「不但」與「反而」之間距離15詞以內，且「反而」與「會」距離2詞以內的句子。
- 一個詞可能被斷詞隔開的查詢：在詞彙中插入 -空格NEAR空格 - 「更 NEAR 重要」
- NEAR未標示距離時，距離預設為NEAR/10，「不但 NEAR 反而」即表示查詢「不但」與「反而」之間距離十個詞內的句子。可依需要調高或調低距離 - 「試想 NEAR/8 愛情」，「試想 NEAR/60 愛情」，「試想 NEAR/64 愛情」

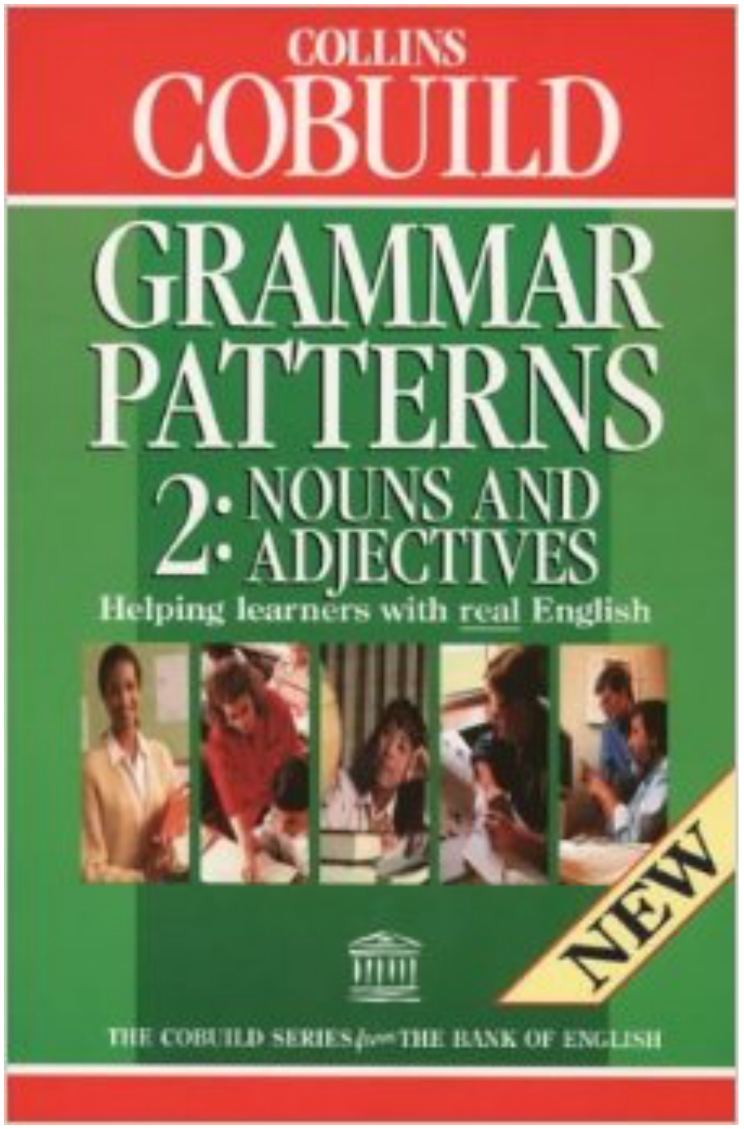
不但 NEAR/15 反而

	Filename	Sentence	Source Id
	9902ES002.txt	她 內心 的 渴求 不但 沒有 隨著 歲月 消減 ， 反而 與日俱增 。	Source
	1001IEPM05010.txt	雖然 這樣 的 舉動 似乎 太 誇張 ， 而且 有點 歧視 的 味道 ， 但是 和 他們 相比 ， 明星 學生 在 教學 資源 上 佔 了 很 大 的 優勢 ， 僅 憑 繁星 計畫 不但 無法 為 偏遠 學生 保留 名額 ， 反而 為 明星 學生 多 開 了 一 條 捷徑 ， 而 藉由 學校 推薦 、 個人 申請 和 指考 統一 分發 ， 明星 高中 每 年 進入 大學 知名 科系 的 也 是 大有人在 ， 代表 僅 憑 考試 成績 進 大學 對 他們 而言 不 是 問題 。	Source
	1041CS03023.txt	」 由 此 可 知 ， 在 真的 見到 了 腳踏車 後 ， 小程 內心 狀態 不但 沒有 好轉 ， 反而 更加 的 迷茫 ， 找 不 到 一 個 既定 的 方向 。	Source

Mandarin Chinese Spelling Check

- End-to-end and multiple grammatical errors types
  - 描瞄准準暑假商機 (sound alike and look alike)
  - 這篇文章很有文學造旨詣 (look alike)
  - 熱淚盈匡眶 (sound alike and look alike)

Future Work


COLLINS
COBUILD
GRAMMAR PATTERNS
2: NOUNS AND ADJECTIVES
Helping learners with real English
NEW
THE COBUILD SERIES from THE BANK OF ENGLISH

- ***GEC*** for Academic Writing
  - Reference Academic Corpus (COCA) + Wikipedia
  - Artificially contrived errors based on data and form of
    - EFCAMDAT
    - Grammar Patterns from Francis, Hunston, Manning (1996,7)
  - Neural Machine Translation
    - sense2vec and collocation2vec
- ***WriteAhead*** for Academic Writing
  - COCA Academic + Wikipedia
  - Handcrafted grammar patterns from FHM 1996,7
    - 9,500 verb patterns, 14,300 noun patterns, 5,700 adjective patterns
  - Automatically derived grammar patterns
- ***Linggle*** for Academic Writing
  - COCA Academic + Wikipedia (English)
  - COCT + Wikipedia (traditional Chinese)