

The Corpus of Historical Singapore English – Practical and Methodological Issues

**SEBASTIAN HOFFMANN
UNIVERSITY OF TRIER**

English as a global language

FROM 16TH CENTURY TO TODAY:

- * Dramatic spread of English – now a global language
- * Establishment of various native varieties (AmE, AusE, NZE, etc.)
- * English as a *lingua franca* – more than 2 billion people are regularly exposed to English (and a good proportion have some competence in English – cf. Crystal 2003)
- * Post-colonial second-language varieties of English (e.g. India, Jamaica, Malaysia)

➡ “New Englishes”

Structural nativisation in the evolution of New Englishes

DEVELOPMENT OF NEW ENGLISHES IN POST-COLONIAL CONTEXTS

- ⇒ "English has been appropriated by its non-European users and changed to reflect their own experiences."

(Mair 2008: 235)

APPROPRIATION ENTAILS "STRUCTURAL NATIVISATION"

- ⇒ [Structural nativisation:] "the emergence of locally characteristic linguistic patterns"

(Schneider 2007: 5f.)

Schneider's (2003, 2007) evolutionary model of the development of New Englishes

- * **Phase I – ‘Foundation’:** In this initial phase, the English language is transported to a new (colonial) territory.
- * **Phase II – ‘Exonormative stabilisation’:** There is a growing number of English settlers/speakers in the new territory (STL strand), but the language standards and norms are still determined by the input variety and are, thus, usually oriented towards British English.
- * **Phase III – ‘Nativisation’:** The English language becomes an integral part of the local linguistic repertoire as there is a steady increase in the number of competent bilingual L2 speakers of English from the indigenous population (IDG strand).

Schneider's (2003, 2007) evolutionary model of the development of New Englishes

- ✱ **Phase IV – ‘Endonormative stabilisation’:** After Independence, English may be retained as a/an (co-)official language and a medium of communication for a more or less wide range of intra-national contexts (e.g. administration and the press, academia and education); in this phase a new variety of English emerges with generally accepted local standards and norms.
- ✱ **Phase V – ‘Differentiation’:** Once a New English variety has become endonormatively stabilised, it may develop a wide range of regional and social dialects.

Investigations of New Englishes

- * Large number of studies, investigating various linguistic features – ranging from phonology to variational pragmatics
- * Many studies are corpus-based
- * ICE: more than 20 varieties of English world-wide, many ESL varieties (cf. Greenbaum 1996)
- * Kolhapur corpus (IndE – data from 1978)
- * Newspaper archives (cf. Schilk 2011, Mukherjee & Hoffmann 2006)

⇒ Historical development?

Investigations of New Englishes

TYPICAL APPROACH:

While in the first step I focused on the evolution of Indian English from a diachronic perspective (including a thorough discussion of the interactions between the STL strand [i.e. settlers] and the IDG strand [i.e. the indigenous population], and of relevant processes of identity construction in the various phases), in the second and third steps **I zoomed in on present-day Indian English from a synchronic perspective.** (Mukherjee, 2007: 181-2; my emphasis)

- ➡ At the moment, investigations of the development of New Englishes are almost exclusively restricted to apparent-time studies.

Investigations of New Englishes

DIFFICULT AREAS IN THE INVESTIGATION OF NEW ENGLISHES:

- * the distinction between learner language phenomena ('mistakes') and genuine nativised regionalisms (particularly in the context of low-frequency features)
- * the role of first-language interference
- * the distinction between fossilised features of earlier stages of the input variety (i.e. typically British English) and real innovations; cf. the concept of 'colonial lag' (e.g. Görlach 1987)
 - ➡ A diachronic corpus of one or several new varieties of English is necessary to confirm Schneider's (2003/2007) evolutionary stages and to investigate the development of New Englishes in real time.

English in Singapore

"By now Singapore has clearly reached phase 4 of the cycle. The country's unique, territory-based, and multicultural identity construction has paved the way for a general acceptance of the local way of speaking English as a symbolic expression of the pride of the Singaporeans in their nation."

(Schneider 2007: 160)

- ✱ Historical development?
- ✱ For SinE, too, no corpora are available to trace (projected) path along Schneider's five phases

'''→ **CORPUS OF HISTORICAL SINGAPORE ENGLISH**

Corpus of Historical Singapore English

TIME-PERIOD: APPROX. 1951 TO 2011, COLLECTED IN 10-YEAR INTERVALS

- * In 1951, English would have been mainly spoken by the STL-strand of the population
- * Exceptions: local elite and small proportion of locals in administration and domestic services

⇒ **SUITABLE STARTING POINT?**

Corpus of Historical Singapore English

CRITERIA FOR INCLUSION:

- * Written texts only (but: oral history interviews?)
- * Standard English – not “Singlish”
- * Authors who live(d) in Singapore and/or texts with target-audience in Singapore?
- * Text published in Singapore
 - ⇒ **RESTRICTION TO (ETHNICALLY) LOCAL AUTHORS IS NEITHER FEASIBLE NOR DESIRABLE**
 - ⇒ **STL-STRAND VS. IDG-STRAND BOTH NEED TO BE REPRESENTED**
 - ⇒ **PROPORTIONS?**

Corpus of Historical Singapore English: Text categories

- * Informative prose (general + academic)
- * Imaginative prose – Fiction
- * Newspapers / periodicals
- * Speeches
- * School yearbooks, essays, exams?
- * Oral history interviews?
- * Unpublished material (letters etc.)?
- * CMC?

A note on copyright/availability

- * Singapore – a corpus compiler’s nightmare
- * Sources: National Archives / National Library / other archives
- * Seeking permission for inclusion of texts into corpus is simply impossible (and too expensive to try)
- * Availability of corpus will have to be restricted:
 - * No distribution
 - * Web-based interface (CQPweb)
 - * Google-like “snippets” as context display

Representativeness

PROBLEM #1: GENRE CONTINUITY

- ✱ Problem common to all diachronic corpora
- ✱ LOB vs. FLOB; B-LOB (1931) & LOB 1901

Representativeness

PROBLEM #2: RADICAL SOCIETAL CHANGES

- * Proportion of speakers of English (within IDG-strand and vs. native speakers of English)
- * Language competence of readers
- * Target audience of texts published in Singapore

EXAMPLE:

Proportion of textbooks and education-related texts is greater in earlier periods

➡ How should this be reflected in corpus compilation?

Representativeness

PROBLEM #2: RADICAL SOCIETAL CHANGES

- * Proportion of speakers of English (within IDG-strand and vs. native speakers of English)
 - * Language competence of readers
 - * Target audience of texts published in Singapore
- ⇒ requires disentangling the "normal" development of genres from the growing acquisition/appropriation of the range of language contexts and uses in a non-native/second-language context.

Corpus of Historical Singapore English: Informative Prose

Home

Singapore National Bibliography

Listing of print and non-print library materials published in Singapore and deposited with the Singapore National Library, under the National Library Board Act, 1995 by the publishers. It reflects the publishing output of Singapore and provides a useful finding aid in searching for materials published in Singapore.

History and geography

900	Geography & history
910	Geography & travel
920	Biography, genealogy, insignia
930	History of the ancient world
940	General history of Europe
950	General history of Asia (Far East)
960	General history of Africa
970	General history of North America
980	General history of South America
990	General history of other areas

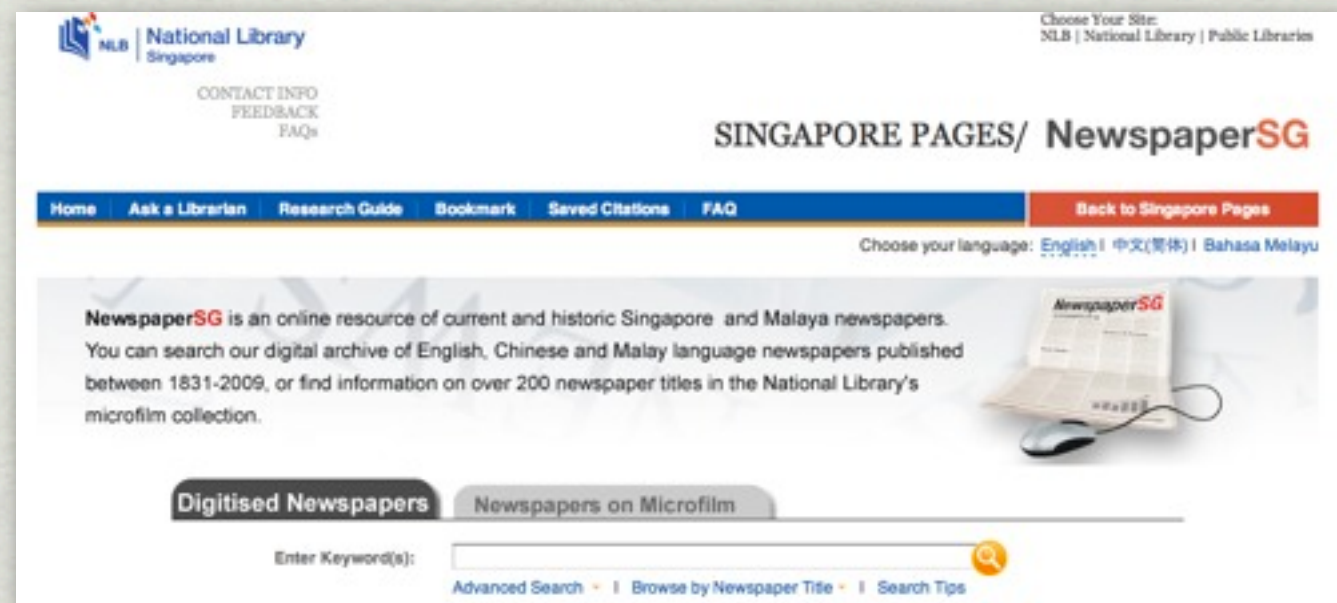
- ✱ Dewey Decimal Classification
- ✱ So far, no distinction between learned and general prose

Newspaper data

TABLOID / BROADSHEET NEWSPAPERS

- ✱ *Straits Times*: general daily broadsheet paper, founded is 1845
- ✱ *The New Paper*: tabloid newspaper founded in 1987
- ✱ No availability of earlier tabloids (?)

Digitised resources
are available:



The screenshot shows the NewspaperSG website interface. At the top left is the National Library Singapore logo. To the right, it says "Choose Your Site: NLB | National Library | Public Libraries". Below the logo are links for "CONTACT INFO", "FEEDBACK", and "FAQs". The main header reads "SINGAPORE PAGES/ NewspaperSG". A navigation bar includes "Home", "Ask a Librarian", "Research Guide", "Bookmark", "Saved Citations", "FAQ", and "Back to Singapore Pages". Below this, it says "Choose your language: English | 中文(简体) | Bahasa Melayu". The main content area features a description: "NewspaperSG is an online resource of current and historic Singapore and Malaya newspapers. You can search our digital archive of English, Chinese and Malay language newspapers published between 1831-2009, or find information on over 200 newspaper titles in the National Library's microfilm collection." To the right is an image of a newspaper on a laptop. At the bottom, there are two tabs: "Digitised Newspapers" (selected) and "Newspapers on Microfilm". Below the tabs is a search bar with the text "Enter Keyword(s):" and a search button. At the very bottom, there are links for "Advanced Search", "Browse by Newspaper Title", and "Search Tips".

Newspaper data

TABLOID / BROADSHEET NEWSPAPERS

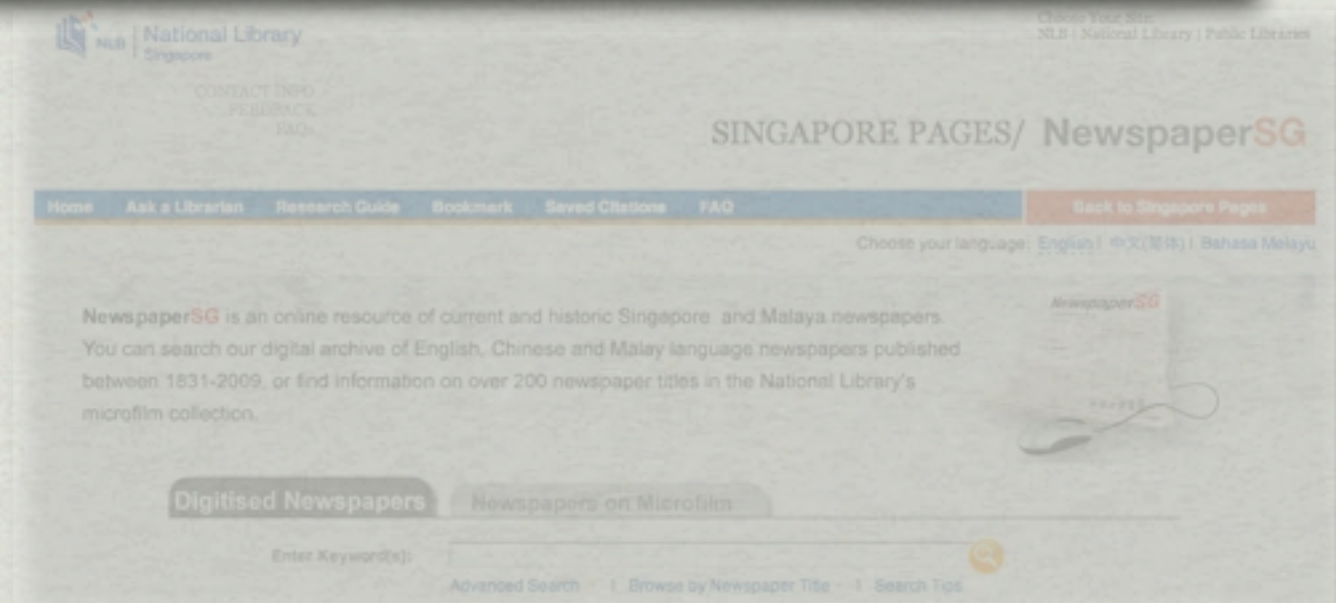
* *Straits Times*: general daily broadsheet paper,

NewspaperSG is an online resource of current and historic Singapore and Malaya newspapers.

You can search our digital archive of English, Chinese and Malay language newspapers published between 1831-2009, or find information on over 200 newspaper titles in the National Library's microfilm collection.

Digitised resources are available:

HOWEVER...



Newspaper data

TABLOID / BROADSHEET NEWSPAPERS

- ✱ *Straits Times & The New Paper* data from 1991 onwards is available from Singapore Press Holdings at reasonable cost
- ✱ Pre-1989 data is restricted to online-search facility run by National Library Board

Newspaper data

MESSAGE FROM THE VERY TOP OF THE ORGANISATION:

My name is Ngian and I am the Director of the National Library of Singapore.

We have reviewed the request from Prof Sebastian Hoffmann and regret to inform you that we are unable to accede to this.

The NLB's newspaperSG service is available to anyone wishing to access it and we welcome Prof Sebastian Hoffmann to use it for his research at his own convenience.

Newspaper data

TABLOID / BROADSHEET NEWSPAPERS

- ✱ *Straits Times & The New Paper* data from 1989 onwards is available from Singapore Press Holdings at reasonable cost
- ✱ Pre-1989 data is restricted to online-search facility run by National Library Board
- ✱ Retrieval fee imposed (on top of licensing fees): S\$10,000
- ✱ New approach: improving the OCR quality in return for free access to the data

Examples of OCR errors in the *Straits Times* data

The patrol reported six of its **s?ven** members were wounded,

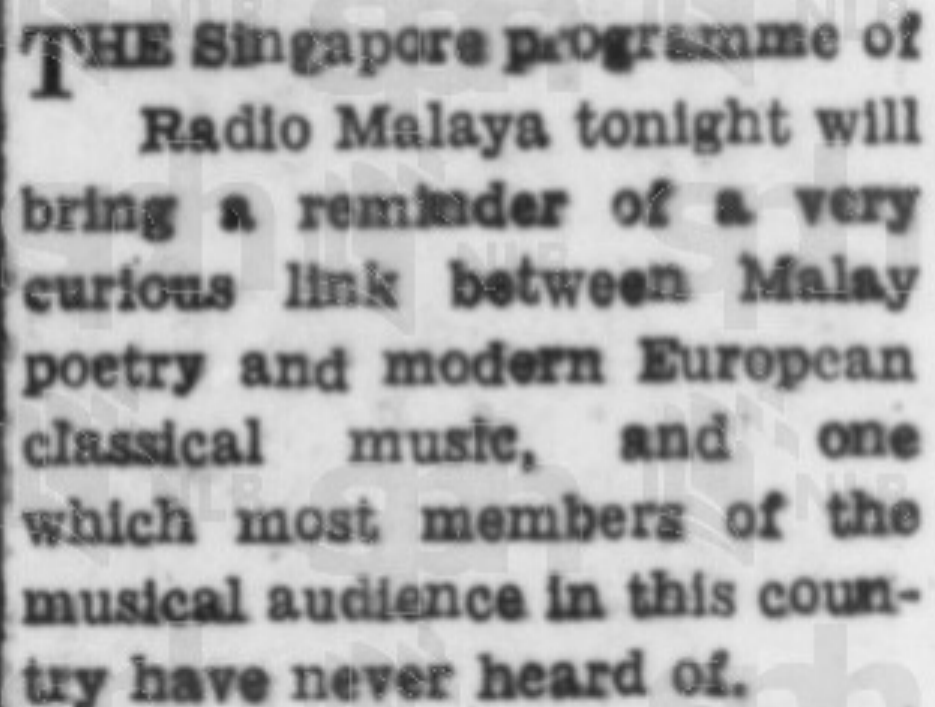
An **otlicial** decision on new financial arrangements may be made this week

Rooftop hunt **lor** a man **WATCiED** by a large crowd

Natidnallst guerillas on the night of Sept. 1 captured two Russian advisers **m** the outskirts of Swatow.

THE **Skicaport pt^frionae** of Radio Malaya tonight will bring a **naifeuUr** of very **curlotw** link **b«tw«««n Malsiy** poetry and modern European classical music, and **on*** which most members of the musical audience **m** this country **°»ye** never heard of.

Newspaper data



THE Singapore programme of Radio Malaya tonight will bring a reminder of a very curious link between Malay poetry and modern European classical music, and one which most members of the musical audience in this country have never heard of.

THE **Skicaport pt^frionae** of Radio Malaya tonight will bring a **naifeuUr** of very **curlotw** link **b«tw«««n Malsiy** poetry and modern European classical music, and **on*** which most members of the musical audience **m** this country **°»ye** never heard of.

Error rates

- * Obviously, older data contains more problems
- * Different text types/genres display different error rates (advertising with its more varied typographic conventions is particularly bad...)



DETAILED ERROR RATES HAVE NOT YET BEEN CALCULATED...

Unsupervised OCR-error correction

- * Previous implementations exist, discussed and described in a sizable number of scholarly articles
- * As far as I am aware, no ready-made off-the-shelf (and commercial) solutions are available
- * Recent diploma thesis:

Niklas, Kai (2010) *Unsupervised Post-Correction of OCR Errors*. Unpublished diploma thesis, Hannover University, Germany. Available at http://www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf.

OCR Error Classification

Non-word error:

A non-word error occurs if the original word is not contained in any dictionary (e.g. *tho* instead of *the*). This is the type of error that spell checkers would attempt to correct (and might do so successfully).

Real-word error:

A real-word error occurs if the original word is correctly spelled but incorrectly used (e.g. *peace of paper* instead of *piece of paper*). This type of error is only correctable using context information.

⇒ **NOT SUFFICIENT FOR OCR ERRORS**

Segmentation errors. Different line, word or character spacings lead to misrecognitions of white-spaces, causing segmentation errors (e.g. *thisis* instead of *this is* or *depa rtmen t* instead of *department*).

Hyphenation errors. Tokens are split up at line breaks if they are too long, which increase the number of segmentation errors (e.g. *de- partment*).

Misrecognition of characters. Dirt and font-variations prevent an accurate recognition of characters which induce wrong recognitions of words (e.g. *souiid* instead of *sound* or *£-Bi1rd#!* instead of *Bird*).

Punctuation errors. Dirt causes misrecognitions of punctuation characters. This means points, commas, etc. occur more often in wrong places with missing or extra white-spaces etc.

Case sensitivity. Due to font variations, upper and lower case characters can be mixed up (e.g. *BrItaIn* or *BRITAIN*).

Changed word meaning. Misrecognized characters can lead to new words which are often wrong in context but spelled correctly (e.g. *mad* instead of *sad*).

Strategies for correcting non-word errors

1. Check whether word is in dictionary/frequency list
2. If not: Retrieve list of similar words from dictionary/frequency list as potential corrections (correction proposals)

Measures to establish the similarity of words:

- * Levenshtein-Distance: Minimum number of insertions, deletions and substitutions of characters which are necessary in order to transform a string x into y
- * n -grams on character level: A character n -gram is a subsequence of n characters of a word. The fraction of n -grams which both words have in common and unique n -grams which both produce, can be used as similarity measure.

Strategies for correcting non-word errors

1. Check whether word is in dictionary/frequency list
2. If not: Retrieve list of similar words from dictionary/frequency list as potential corrections (correction proposals)

Measures to establish the similarity of words:

- * OCR-Key (Niklas 2010): Computation of a key for each word based on the similarity of characters in terms of structure

mni

original

nni

damaged

lllll

structure

Strategies for correcting non-word errors

Class	Member	Cardinality
i	f, i, j, k, l, r, t, B, D, E, F, I, J, K, L, P, R, T, 1, !	1
i	n, h, u, H, N, U	2
i	m, M	3
o	a, b, d, g, o, p, q, O, Q, 6, 9, 0	1
c	e, c, C, G	1
v	v, x, y, V, Y, X	1
v	w, W	2
s	s, S, 5	1
z	z, Z	1
a	A	1

Table 3.2: Character Classification by Similarity

Strategies for correcting non-word errors

saturday → $\underbrace{s}_{s1} \underbrace{a}_{o1} \underbrace{tur}_{i4} \underbrace{da}_{o2} \underbrace{v}_{v1}$ → *s1o1i4o2v1*

(NIKLAS 2010: 29)

⇒ **STRUCTURALLY SIMILAR WORDS WILL HAVE VERY SIMILAR (OR EQUAL) OCR-KEYS**

But:

Structurally similar words may in fact be very different words:

word	OCR-Key
<i>minimum</i>	i15
<i>untruthful</i>	i15

⇒ **APPLY LEVENSHTAIN-DISTANCE MEASUREMENT TO DISCARD UNLIKELY CANDIDATES**

Additional issues: First words in articles

Initials:

**Eng Tong makes
record lift**
KOH ENG TONG, Malaya's
featherweight represen-
tative to the British Empire
Games in Auckland, broke
the Singapore feather-
weight lifting record of

Eng Tong makes record lift |fOH
ENG TONG, Malaya's
featherweight representative to
the British Empire Games in
Auckland, broke the Singapore
featherweight lifting record of

Free Press Staff Reporter
CLIPPIES may soon appear
in Singapore now that
the application of the Sin-
gapore-Johore Express Ltd.

M.IPPIKS may soon appear in
Singapore now that the
application of the Singapore-
Johore Express Ltd.

→ MAY POTENTIALLY BE CORRECTED IF FINAL LETTERS ARE CORRECT

“Local” words

<918> <KELANTAN OLE BOYS' DINNER [Articles]> <The Straits Times, 10 January 1950, Page 4> **KELANTAN OLE BOYS' DINNER** From Oar Own Correspondent **KOTA BAHRU**, Mon. About 100 old boys and guests, including the Sultan, the **Tungku Mahkota**, the **Mentri Besar**, the British Adviser, the State Secretary **Inche Mahmood Mahyiddin** (former Superintendent **oi** Education) and Mr. **Wonp Quek** Boon (Headmaster) attended the <http://newspapers.nl>.

A considerable proportion of unrecognized words are in fact perfectly correct (Kota Bahru, Tungku, Mentri Besar etc.)

⇒ **RELEVANT DICTIONARY/FREQUENCY LIST IS REQUIRED TO AVOID “CORRECTING” SUCH WORDS**

Dictionaries/frequency lists

Available data:

- * *Straits Times* data from 1989 onwards, available through SPH (but: licensing?)
- * The New York Times Annotated Corpus – over 1 billion words (1.8 million articles) of AmE newspaper data, 1987-2007
- * Google Books n-grams (500 billion words)
- * ICE-Singapore, 1 million words, of which 400,000 written
- * British National Corpus (BNC) – approx. 100 million words of British English (spoken and written), up to year 1993

Format of XML-files

```
<String ID="P1_ST00052" HPOS="943" VPOS="1026" WIDTH="61" HEIGHT="24"  
CONTENT="tore" WC="0.99" CC="1000"/>
```

```
<SP ID="P1_SP00041" HPOS="1004" VPOS="1056" WIDTH="17"/>
```

```
<String ID="P1_ST00053" HPOS="1021" VPOS="1023" WIDTH="75" HEIGHT="32"  
CONTENT="huge" WC="0.98" CC="0010"/>
```

```
<SP ID="P1_SP00042" HPOS="1096" VPOS="1056" WIDTH="16"/>
```

```
<String ID="P1_ST00054" HPOS="1112" VPOS="1031" WIDTH="69" HEIGHT="24"  
CONTENT="gnps" CORR="guns" WC="0.56" CC="9701"/>
```

```
<SP ID="P1_SP00043" HPOS="1181" VPOS="1056" WIDTH="14"/>
```

```
<String ID="P1_ST00055" HPOS="1195" VPOS="1022" WIDTH="30" HEIGHT="28"  
CONTENT="in" WC="0.96" CC="01"/>
```

```
<SP ID="P1_SP00044" HPOS="1225" VPOS="1056" WIDTH="19"/>
```

```
<String ID="P1_ST00056" HPOS="1244" VPOS="1023" WIDTH="48" HEIGHT="27"  
CONTENT="the" WC="1.00" CC="000"/>
```


Format of XML-files

```
<String ID="P1_ST00052" HPOS="943" VPOS="1026" WIDTH="61" HEIGHT="24"  
CONTENT="tore" WC="0.99" CC="1000"/>
```

```
<SP ID="P1_SP00041" HPOS="1004" VPOS="1056" WIDTH="17"/>
```

```
<String ID="P1_ST00053" HPOS="1021" VPOS="1023" WIDTH="75" HEIGHT="32"  
CONTENT="huge" WC="0.98" CC="0010"/>
```

```
<SP ID="P1_SP00042" HPOS="1096" VPOS="1056" WIDTH="16"/>
```

```
<String ID="P1_ST00054" HPOS="1112" VPOS="1031" WIDTH="69" HEIGHT="24"  
CONTENT="gnps" CORR="guns" WC="0.56" CC="9701"/>
```

```
<SP ID="P1_SP00043" HPOS="1181" VPOS="1056" WIDTH="14"/>
```

```
<String ID="P1_ST00055" HPOS="1195" VPOS="1022" WIDTH="30" HEIGHT="28"  
CONTENT="in" WC="0.96" CC="01"/>
```

```
<SP ID="P1_SP00044" HPOS="1225" VPOS="1056" WIDTH="19"/>
```

```
<String ID="P1_ST00056" HPOS="1244" VPOS="1023" WIDTH="48" HEIGHT="27"  
CONTENT="the" WC="1.00" CC="000"/>
```


Format of XML-files

```
<String ID="P1_ST00052" HPOS="943" VPOS="1026" WIDTH="61" HEIGHT="24"  
CONTENT="tore" WC="0.99" CC="1000"/>
```

```
<SP ID="P1_SP00041" HPOS="1004" VPOS="1056" WIDTH="17"/>
```

```
<String ID="P1_ST00053" HPOS="1021" VPOS="1023" WIDTH="75" HEIGHT="32"  
CONTENT="huge" WC="0.98" CC="0010"/>
```

```
<SP ID="P1_SP00042" HPOS="1096" VPOS="1056" WIDTH="16"/>
```

```
<String ID="P1_ST00054" HPOS="1112" VPOS="1031" WIDTH="69" HEIGHT="24"  
CONTENT="gnps" CORR="guns" WC="0.56" CC="9701"/>
```

```
<SP ID="P1_SP00043" HPOS="1181" VPOS="1056" WIDTH="14"/>
```

```
<String ID="P1_ST00055" HPOS="1195" VPOS="1022" WIDTH="30" HEIGHT="28"  
CONTENT="in" WC="0.96" CC="01"/>
```

```
<SP ID="P1_SP00044" HPOS="1225" VPOS="1056" WIDTH="19"/>
```

```
<String ID="P1_ST00056" HPOS="1244" VPOS="1023" WIDTH="48" HEIGHT="27"  
CONTENT="the" WC="1.00" CC="000"/>
```


Comparison with n-gram data

	Straits Times	NYT
huge guns	0	9
huge gaps	3	101
huge guns in	0	0
huge gaps in	2	60

Named entities

The rugged Sobaek Mountain ranges

→ Corrected to: The rugged **Sebarok** Mountain ranges

Sobaek Mountain is not in n-gram data

But it is in the list of Wikipedia titles!

But so is every known asteroid (and similarly unhelpful data: >13 million entries)....

Strategies for correcting real-word errors

- * n -grams on word level (i.e. sub-sequences of n words of a sentence or text and their frequencies)
- * Collocation information (measures and evaluates likelihood of two words co-occurring within a certain window in a text as opposed to other potential – and structurally similar – words)
- * Part-of speech information (POS): include the likelihood of POS-tag sequences as additional criterion

⇒ APPLYING SUCH STRATEGIES MAY RESULT IN REPLACEMENT OF PERFECTLY CORRECT WORD THAT IS FOUND IN A RARE COMBINATION WITH ANOTHER (SET OF) WORD(S)

Straits Times post 1990: SPH vs. NLB data

NLB version:

The Straits Times, 22 January 2000, Page 76: FIVE FROGS IN A POND FIVE frogs are sitting on a log in a pond. One decides to **jump** into the water. How many are **(left** sitting on the log? The riddle was posed to 300 grassroots leaders at a Singapore 21 organised by the Ang Mo Kio-Cheng San Community...

SPH version:

decides to **jump** into the water. How many are **left** sitting on

⇒ **DATA EXISTS IN NON-OCR-ED, ELECTRONIC FORMAT!**

⇒ **COULD BE USED AS TRAINING DATA...**

References

- Görlach, Manfred (1987): “Colonial lag? The alleged conservative character of American English and other “colonial” varieties”, *English World-Wide* 8, 41-60.
- Greenbaum, S., ed. (1996): *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon.
- Mair, Christian (2008): *English Linguistics: An Introduction*. Tübingen: Gunter Narr.
- Mukherjee, Joybrato. (2007): “Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium.” *Journal of English Linguistics* 35: 157–87.
- Mukherjee, J. & S. Hoffmann (2006): “Describing verb-complementational profiles of New Englishes: a pilot study of Indian English.” *English World-Wide* 27(2), 147–173.
- Niklas, Kai (2010) *Unsupervised Post-Correction of OCR Errors*. Unpublished diploma thesis, Hannover University, Germany. Available at <http://www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf>.
- Schilk, Marco (2011): *Structural Nativization in Indian English Lexicogrammar*. Amsterdam & Philadelphia: John Benjamins.
- Schneider, Edgar W. (2003): “The dynamics of New Englishes: From identity construction to dialect birth. *Language* 79(2), 233–281.
- Schneider, Edgar W. (2007): *Postcolonial English. Varieties around the world*. Cambridge: Cambridge University Press.

**THANK YOU FOR YOUR
ATTENTION**