

# Tracking Sociohistorical Trends in the Use of Roman Letters in Chinese Newswires

Helena Riha

Kirk Baker

Department of Linguistics  
The Ohio State University

AACL 2008

March 15, 2008

Provo, Utah

Chinese has a long history of borrowing foreign terms

e.g., Saraśvati 'Hindu goddess of arts'  
莎赖婆底 *shā-lài-pó-dǐ*

Sanskrit (6<sup>th</sup> Century)

Typically adapted via phonological transliteration

Chinese orthography moderates the conversion to Chinese morpho-phonology

## More recent examples of phonological adaptation

Michael Jackson	迈克尔·杰克逊	<i>màikèěr · jiékèxùn</i>
sofa	沙发	<i>shāfā</i>
chocolate	巧克力	<i>qiǎokèlì</i>

Loan translation and semantic adaptation have been used extensively since the 20<sup>th</sup> century

loan translation: basketball 篮球 *lánqiú*  
*lit. basket-ball*

semantic translation: typewriter 打字机 *dǎzìjī*  
*lit. hit-character-machine*

Combinations of phonological adaptation and semantic/loan translation also occur

hacker	黑客	<i>hēi-kè</i>	<i>lit. wicked-visitor</i>
hippie	嬉皮士	<i>xī-pí-shì</i>	<i>lit. grin-cheekily-person</i>

More recently, some foreign terms are being borrowed “as-is” (mainly from English)

## Results in mixed-script texts

APEC 记者招待会后，我约了 STV 的记者和一群 MBA、MPA 研究生朋友，讨论中国加入 WTO 后 IT 业对 GDP 的影响。读 MBA 的张小姐本来想去 .COM 当 CEO，但觉得 IT 业风险大...随后大家相约关掉 BP 机，不上 Internet 的 QQ 和 BBS 聊天，而是去了 KTV 唱卡拉 OK。

from Zhang (2005): People's Daily Newspaper, 4/20/2005

Mixed-script writing is interesting because Chinese and English writing systems denote different linguistic units

Chinese orthography is morpho-syllabic

- Chinese characters are typically monosyllabic, monomorphemic
- Written without indication of word boundaries

English orthography is lexico-phonemic

- Words are delineated with white space
- Individual characters typically convey pronunciation

Leads to questions about how this mismatch is dealt with

How do Roman letters function in Chinese?

- Phonemic (as in English)?
- Morphological (as in Chinese)?

Claim: early Roman letter borrowings were implicitly sinicized

Roman letters primarily function morpho-syllabically by analogy to Chinese characters

English-like borrowings occurred secondarily and are tied to increasing familiarity with English language

Evidence: trends in Roman letter usage in the Chinese Gigaword Corpus

## Chinese Gigaword Third Edition

(Graff 2007); LDC2007T38

16 years of newswire (1991-2006) from two sections of the corpus

Xinhua: 1.1M articles; 545M characters

Central News Agency: 1.9M articles; 952M characters

Represent two politically and socially distinct Chinese societies

People's Republic of China (Xinhua)

Taiwan (Chinese News Agency)

*[Hereafter PRC and Taiwan]*

[Image source: <http://www.gio.gov.tw/taiwan-website/5-gp/rocprc/map.htm>]

Officially, Taiwan is a Chinese province

Unofficially, Taiwan has been politically and socially independent since end of the Chinese civil war in 1949



[Image source: <http://www.gio.gov.tw/taiwan-website/5-gp/rocprc/map.htm>]

Taiwan has had close ties with the West since the 1950's

The PRC has opened up to the West primarily since economic policy shift in the 1980's encouraging international trade and investment



[Image source: <http://www.gio.gov.tw/taiwan-website/5-gp/rocprc/map.htm>]

Political and economic history has led to differences in respective levels of Chinese-English language contact

‘Internationalized’ Taiwan has had longer and more intense contact

‘Internationalizing’ PRC has had shorter and less intense contact



[Image source: <http://www.gio.gov.tw/taiwan-website/5-gp/rocprc/map.htm>]

We expect differences in the use of Roman letters in the two societies based on their relative familiarity with English



## Chinese Gigaword Third Edition

(Graff 2007); LDC2007T38

### gzip'd, UTF8 encoded text files

In UTF8, English alphabet occupies a fixed code range

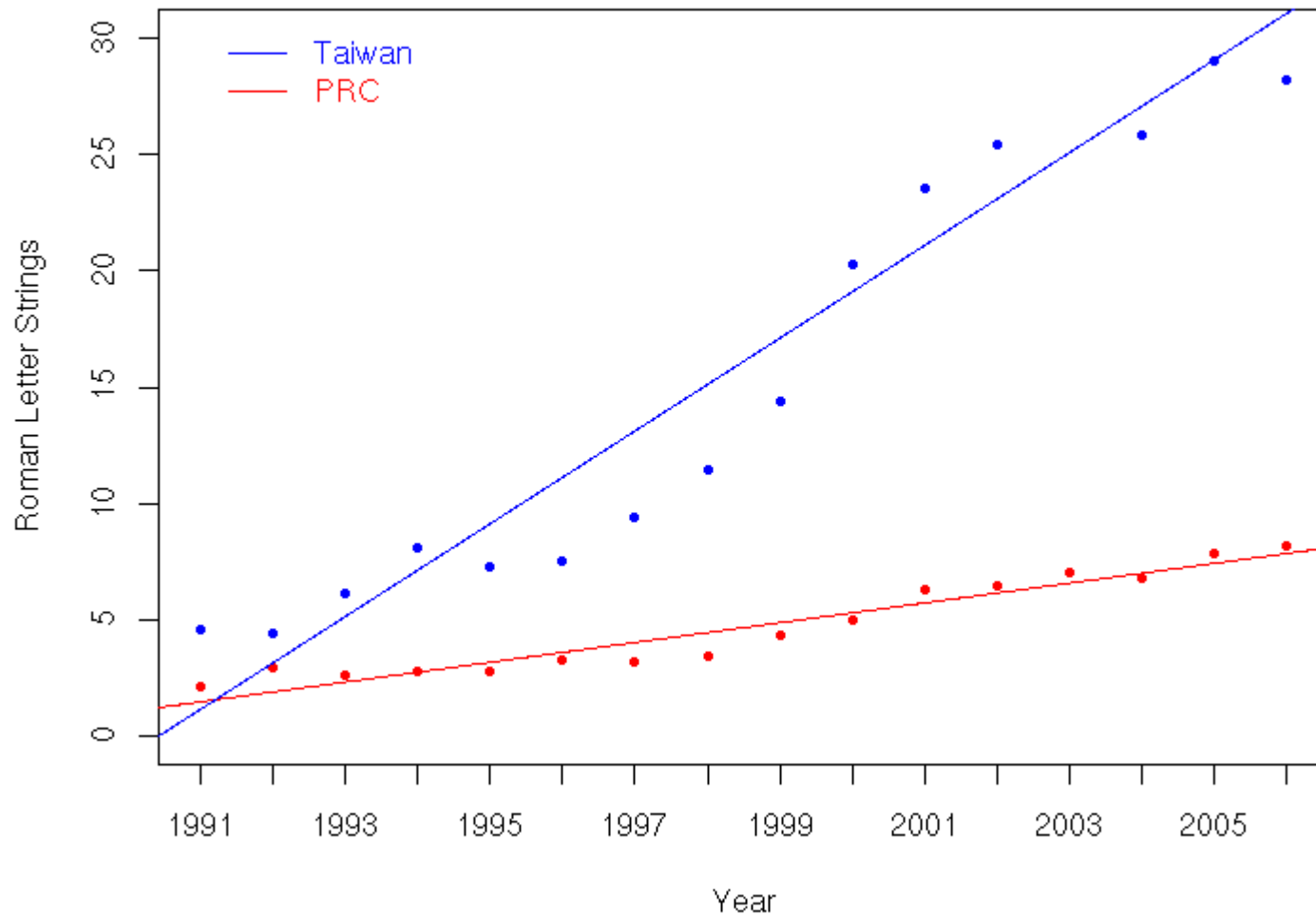
Easy to extract Roman letter strings using regular expressions

Used Python and associated re, gzip, codecs modules to extract all occurrences of Roman letter strings in the PRC and Taiwan sections of the corpus

These data form the basis of subsequent analyses

## General trends in Roman letter usage

Number of Roman Letter Strings per 10k Chinese Characters

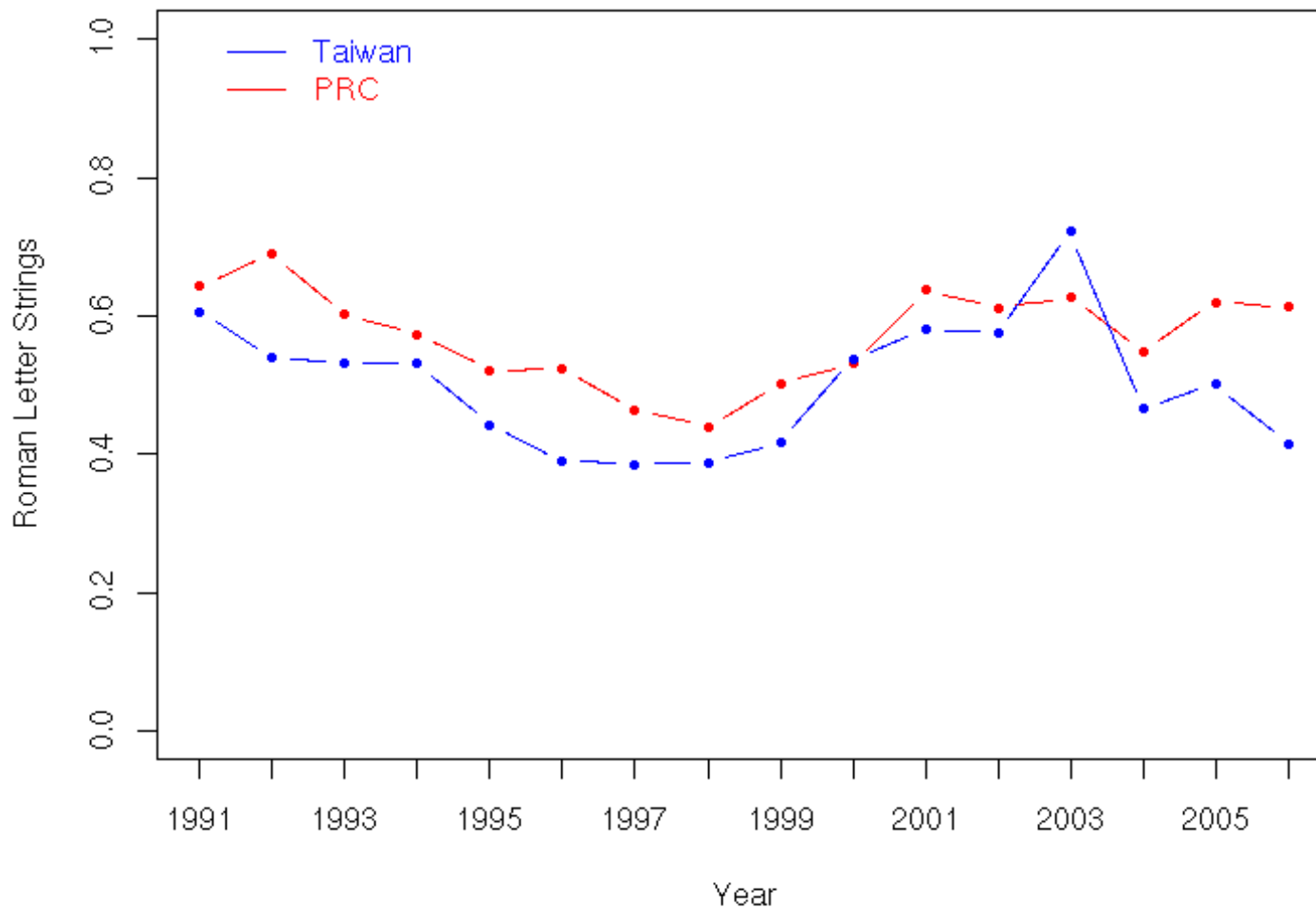


Similarities  
increase over time

highly correlated

## General trends in Roman letter usage

Scaled and rotated

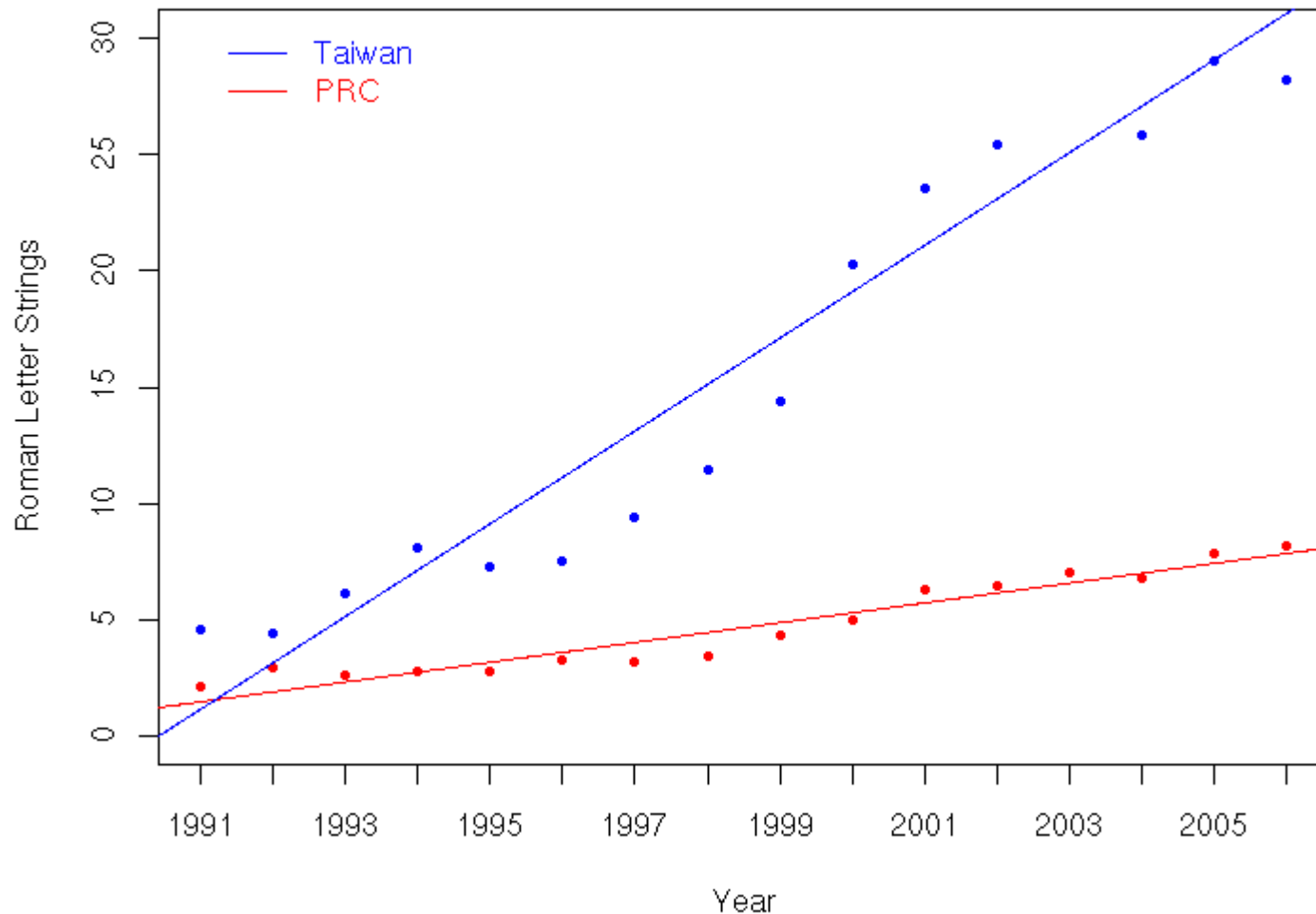


Similarities  
highly correlated

$$r^2 = 0.95$$

## General trends in Roman letter usage

Number of Roman Letter Strings per 10k Chinese Characters



### Differences

overall, more frequent in Taiwan

increasing faster in Taiwan

## Interpretation of Results

The use of Roman letter strings has increased in both societies as the use of English and familiarity with English have become more widespread

Taiwan had earlier and more intense contact with English, leading to more familiarity with English and a higher societal level of fluency than the PRC

*use of roman letters there is greater overall and has had a faster rate of growth*

Taiwan's greater political and social openness and its position as an international trading power most likely contribute to the steep rise in the use of Roman letters

## Interpretation of Results

Broad use of English in the PRC began only in the 1980s and the societal level of fluency is not as high as in Taiwan  
*use of roman letters is lower overall and rate of growth is slower*

The PRC's "controlled" openness to international influences and relatively closed political system may be a factor limiting the use of roman letters

The PRC's official language policy promoting Mandarin as the national language limits the use of foreign Roman letter items in official publications.

## Representative types of Roman letter strings

Proper names: World Trade Organization, Bruce Lee

Hybrid compounds: X 光 ‘ X-ray’, ATM 机 ‘ ATM machine’

Other hybrid strings: 乔治• W• 布希 ‘ George W Bush’

Phrases: Thank You, do your best

Initialisms/Acronyms: WTO, NBA, CPA

## Targeted analysis: initialisms vs words

### Initialisms

letters are generally pronounced as individual syllables

letters may be “morphemic” in that each one can contribute a distinct semantic component to the term

acronymic use of Roman letters is analogous to use of Chinese characters

## Targeted analysis: initialisms vs words

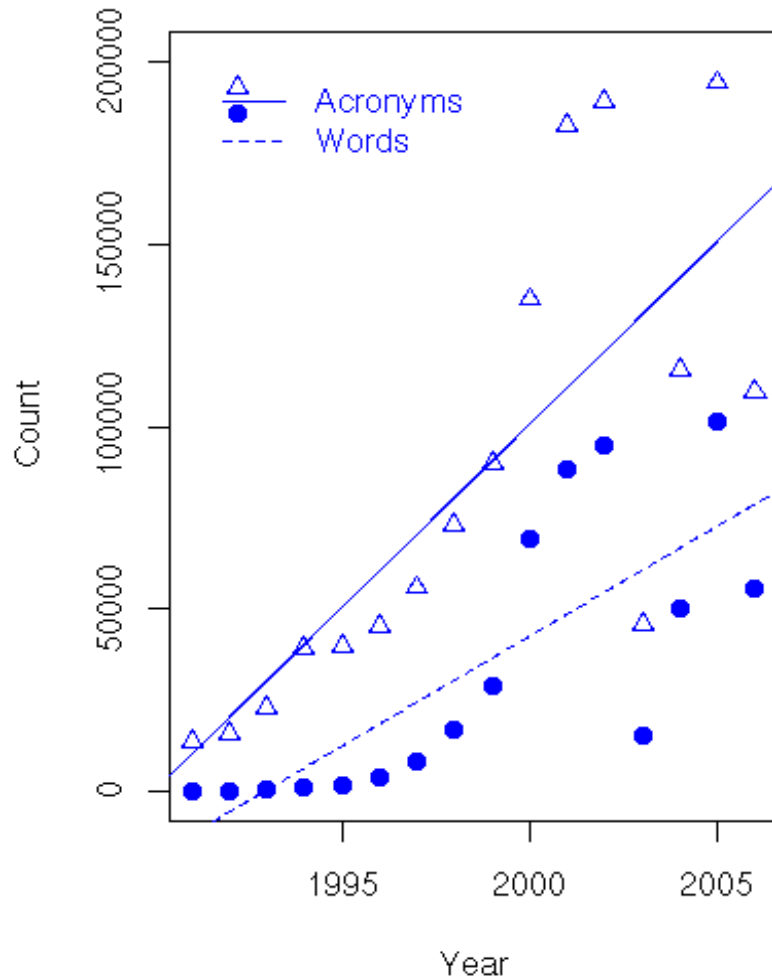
We approximate distinction with regular expressions

initialism  $\approx$  sequence of capital letters

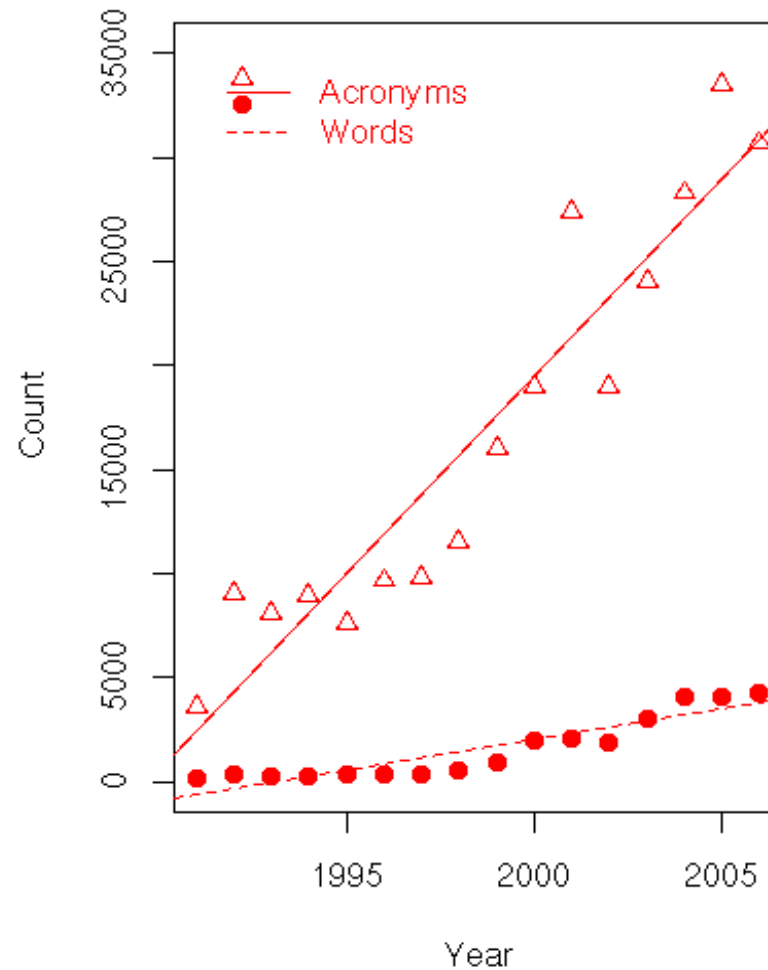
word  $\approx$  sequence of upper and lower case letters

## Targeted analysis: initialisms vs words

Taiwan



PRC



## Targeted analysis: initialisms vs words

### Taiwan

rate of growth of words is nearly the same as for initialisms

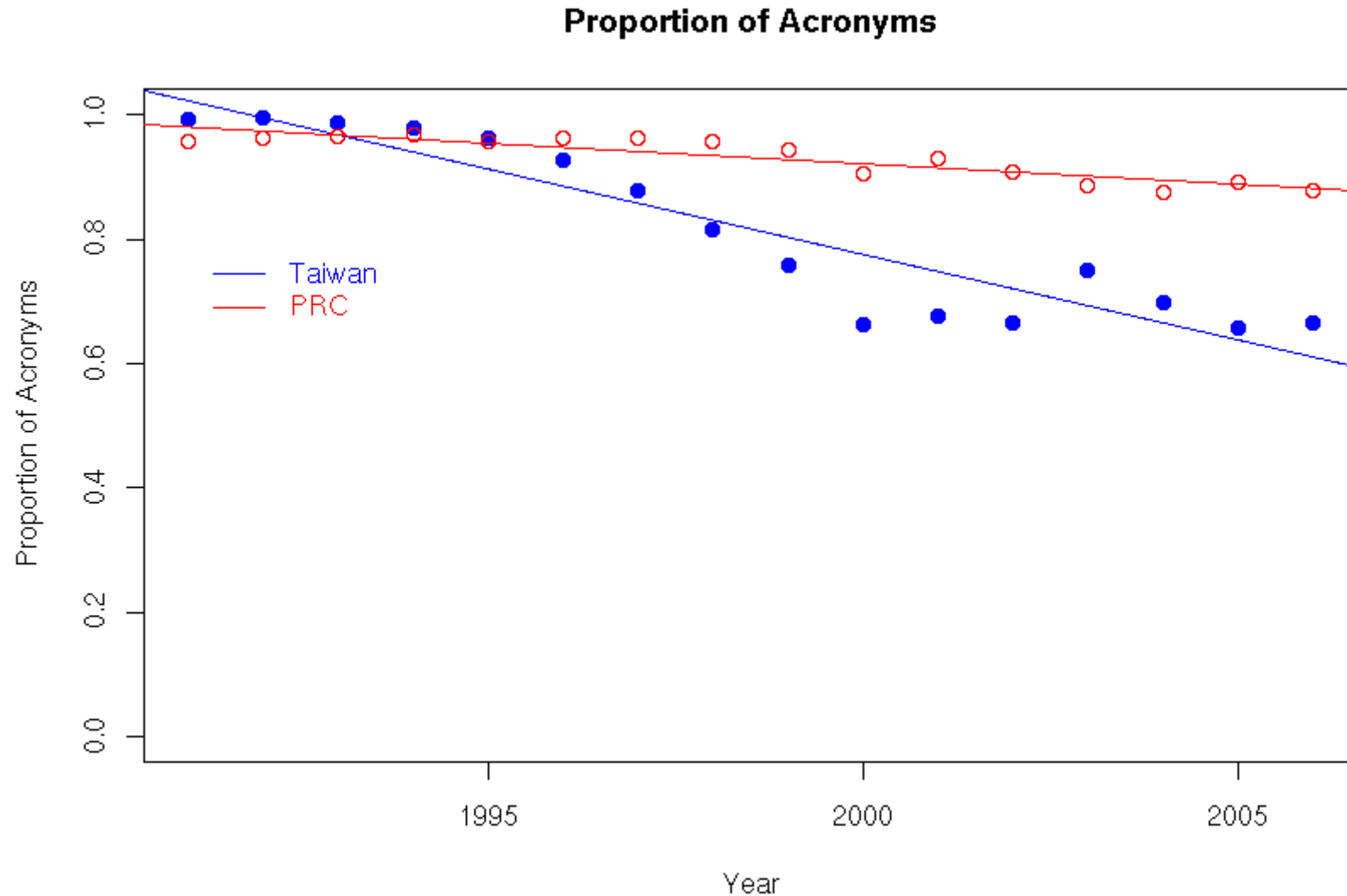
relatively high level of English familiarity

### PRC

nearly all of the increase is due to increased use of initialisms

level of societal familiarity with English is still relatively low, still a preference for using English letters in the manner of Chinese characters

## Targeted analysis: acronyms vs words



## Targeted analysis: initialisms vs words

For both societies, earlier borrowings were overwhelmingly initialisms

Interpret this as evidence for our claim that early Roman letter borrowings were facilitated by analogy to Chinese morpho-syllabic characters

Proportion of initialisms decreases over time for both societies

Interpret this as less reliance on analogy as familiarity with English increases

# Conclusions

The use of Roman letter strings is increasing in Taiwan and PRC

Most of the increase in PRC is due to initialisms

The increase in Taiwan is more evenly spread between words and initialisms

These differences reflect language attitudes and sociohistorical developments in the two societies