Comparing Government Website English Texts Between China and UK+US: Ideas behind, corpus building issues

03/11/2016

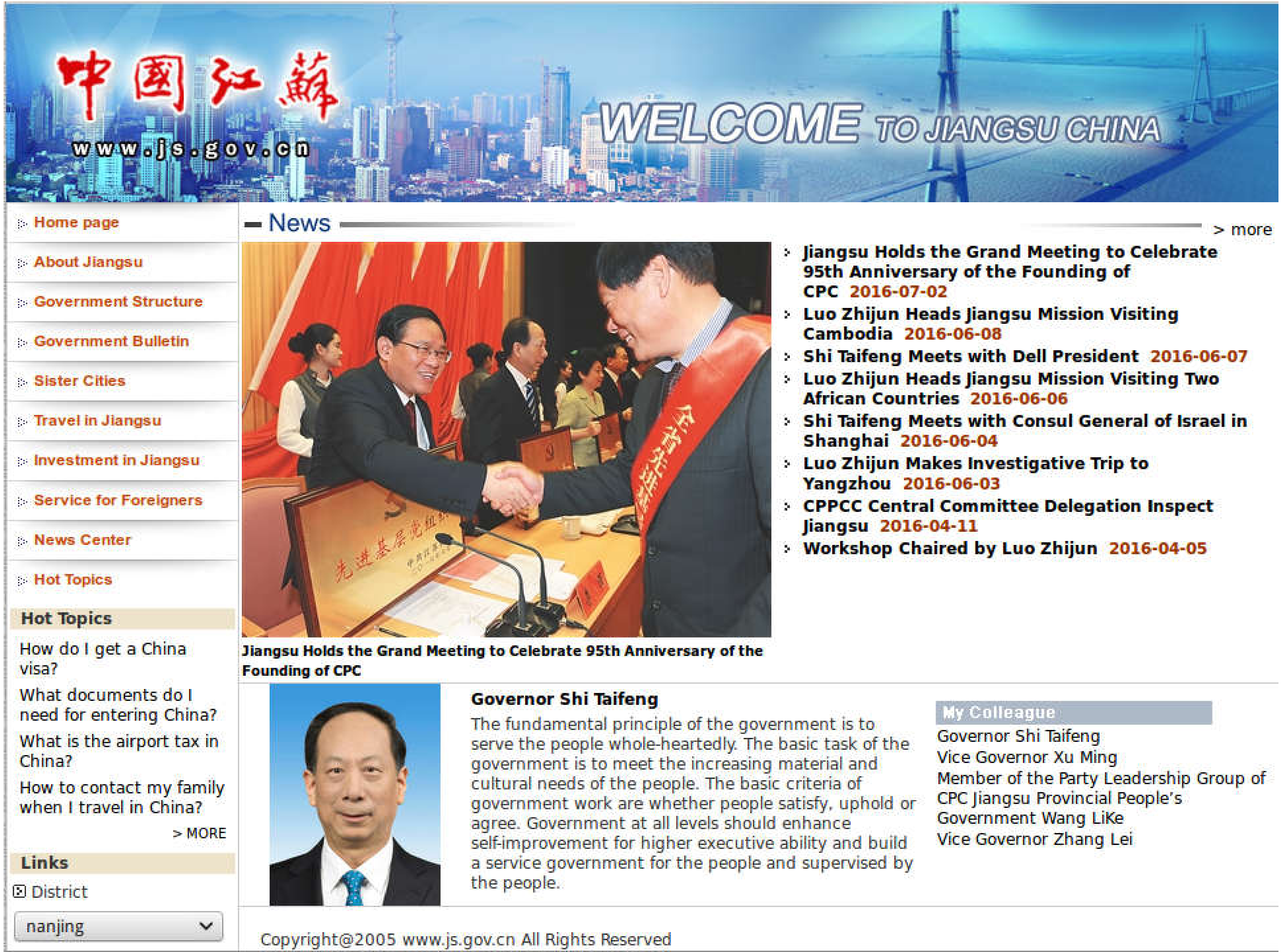
Dr Jiayue Wang / arthur0421@gmail.com

What is meant by “Chinese/UK+US government(al) websites”?

These refer to the websites hosted by central/federal or provincial/state governments, including

- China central government and provincial governments
- UK central government and England, Wales, Scotland, Northern Ireland governments
- US federal government and state governments

Comparability?



News

- **Jiangsu Holds the Grand Meeting to Celebrate 95th Anniversary of the Founding of CPC** 2016-07-02
- **Luo Zhijun Heads Jiangsu Mission Visiting Cambodia** 2016-06-08
- **Shi Taifeng Meets with Dell President** 2016-06-07
- **Luo Zhijun Heads Jiangsu Mission Visiting Two African Countries** 2016-06-06
- **Shi Taifeng Meets with Consul General of Israel in Shanghai** 2016-06-04
- **Luo Zhijun Makes Investigative Trip to Yangzhou** 2016-06-03
- **CPPCC Central Committee Delegation Inspect Jiangsu** 2016-04-11
- **Workshop Chaired by Luo Zhijun** 2016-04-05

Jiangsu Holds the Grand Meeting to Celebrate 95th Anniversary of the Founding of CPC

Hot Topics

How do I get a China visa?

What documents do I need for entering China?

What is the airport tax in China?

How to contact my family when I travel in China?

> MORE

Governor Shi Taifeng

The fundamental principle of the government is to serve the people whole-heartedly. The basic task of the government is to meet the increasing material and cultural needs of the people. The basic criteria of government work are whether people satisfy, uphold or agree. Government at all levels should enhance self-improvement for higher executive ability and build a service government for the people and supervised by the people.

Governor Shi Taifeng
Vice Governor Xu Ming
Member of the Party Leadership Group of CPC Jiangsu Provincial People's Government Wang LiKe
Vice Governor Zhang Lei

A piece of text from that page

Governor Shi Taifeng

The fundamental principle of the government is to serve the people whole-heartedly. The basic task of the government is to meet the increasing material and cultural needs of the people. The basic criteria of government work are whether people satisfy, uphold or agree. Government at all levels should enhance self-improvement for higher executive ability and build a service government for the people and supervised by the people.

General observation

Part of the Chinese websites haven't lived up to what they are designed to.

Language use

- English texts on government websites in China are often said to be 'Chinglish', 'lifeless'.

Page appearance

- The layout, colouring, typefacing etc. of these web pages on these sites largely following the practice of Chinese webpage design.

Therefore, the English texts need improvement in order to achieve better “propoganda effect”.

Research question

Major question

What are the main differences between Chinese government websites and UK/US ones, in the linguistic and textual aspects:

- Typical expressions (keywords, ngrams/clusters)
- Collocates of significant keywords
- Semantic and discourse prosody of certain keywords

Related studies

Most of the related studies were done in the perspectives of communcation, management, security etc. A relatively small part of them discuss the language aspect of the differences. Among those who focus on the English used on government websites, Zhu et al. (2009), Zhu (2010), Wang (2012), Yao (2012), Tian (2013), Zhu (2014), Shao (2014), Ran (2015) are more recent but not based on research.

Yan (2013) conducted a survey of the English versions of the provincial governments' websites.

Methodology

The research is seen as a comparison between English as a foreign language (mostly translated English) and English as a native language.

The research will be a corpus-based, stylistic comparison.

A series of features and dimensions will be investigated to find out the main differences.

Cf:

- Xiao (2010) investigated the differences between translated Chinese (ZCTC, 2001) and native Chinese (LCMC, 1991);
- Chen (2012) compares the use of word clusters in English translated texts and native English texts.

General corpus plan

- Find the URLs of the websites of interest;
- Download the websites;
- Build a corpus that includes 2 subcorpora (Chinese government websites; UK/US government websites)
    - Only English texts;
    - No “boilerplate” parts.

URLs of the Chinse websites (EN version)

Anhui	http://english.ah.gov.cn/
Beijing	http://www.ebeijing.gov.cn/
Central	http://english.gov.cn/
Chongqing	http://en.cq.gov.cn/index.shtml
Fujian	http://www.fujian.gov.cn/english/
Gansu	http://www.gansu.gov.cn/col/col3302/index.html
Guizhou	http://www.eguizhou.gov.cn/
Hainan	http://en.hainan.gov.cn/englishgov/
Hebei	http://www.hebei.gov.cn/english/index.html
Hubei	http://en.hubei.gov.cn/
Hunan	http://www.enghunan.gov.cn/
Jiangsu	http://218.94.123.16:8080/pub/jsgov/JSGOVEN08/index.html
Jiangxi	http://english.jiangxi.gov.cn/
Jilin	http://english.jl.gov.cn/
Shaanxi	http://english.shaanxi.gov.cn/
Shanghai	http://www.shanghai.gov.cn/shanghai/node27118/index.html
Sichuan	http://www.sc.gov.cn/10462/10758/index.shtml

Only 16 out of 30 provicial government websites provide EN editions.

Chongqing gov's EN version found on 27/10/2016

Link to the Zhejiang URL only visible inside flash video

URLs of the US websites

https://www.usa.gov/
http://www.alabama.gov/
http://alaska.gov/
https://az.gov/
http://www.arkansas.gov/
http://www.ca.gov/
https://www.colorado.gov/
http://portal.ct.gov/
https://delaware.gov/
http://www.myflorida.com/
http://georgia.gov/
https://portal.ehawaii.gov/
http://www.idaho.gov/ **
https://www.illinois.gov/
http://www.in.gov/
https://www.iowa.gov/
https://www.kansas.gov/

http://kentucky.gov/
http://louisiana.gov/
http://www.maine.gov/
http://www.maryland.gov/
http://www.mass.gov/
https://www.michigan.gov/
https://mn.gov/
http://www.ms.gov/
http://www.mo.gov/
http://mt.gov/d *
http://www.nebraska.gov/
http://nv.gov/
https://www.nh.gov/
http://www.nj.gov/
http://www.newmexico.gov/
http://www.ny.gov/
http://www.nc.gov/

http://www.nd.gov/
https://ohio.gov/
https://www.ok.gov/
http://www.oregon.gov/
http://www.pa.gov/
http://www.ri.gov/
http://www.sc.gov/ **
http://sd.gov/
https://www.tn.gov/
https://www.texas.gov/
http://www.utah.gov/
http://www.vermont.gov/
https://www.virginia.gov/
http://access.wa.gov/
http://www.wv.gov/
http://www.wisconsin.gov/
http://www.wyo.gov/

For various reasons only 40 were downloaded during a period of 7 months.

* Coudn't download

** Couldn't open even if visited in the UK

URLs of the UK websites

https://www.gov.uk/

http://www.gov.scot/

http://gov.wales/?lang=en

https://www.northernireland.gov.uk/

“England does not have a devolved parliament or regional assemblies, outside Greater London.” (https://en.wikipedia.org/wiki/Local_government_in_England)

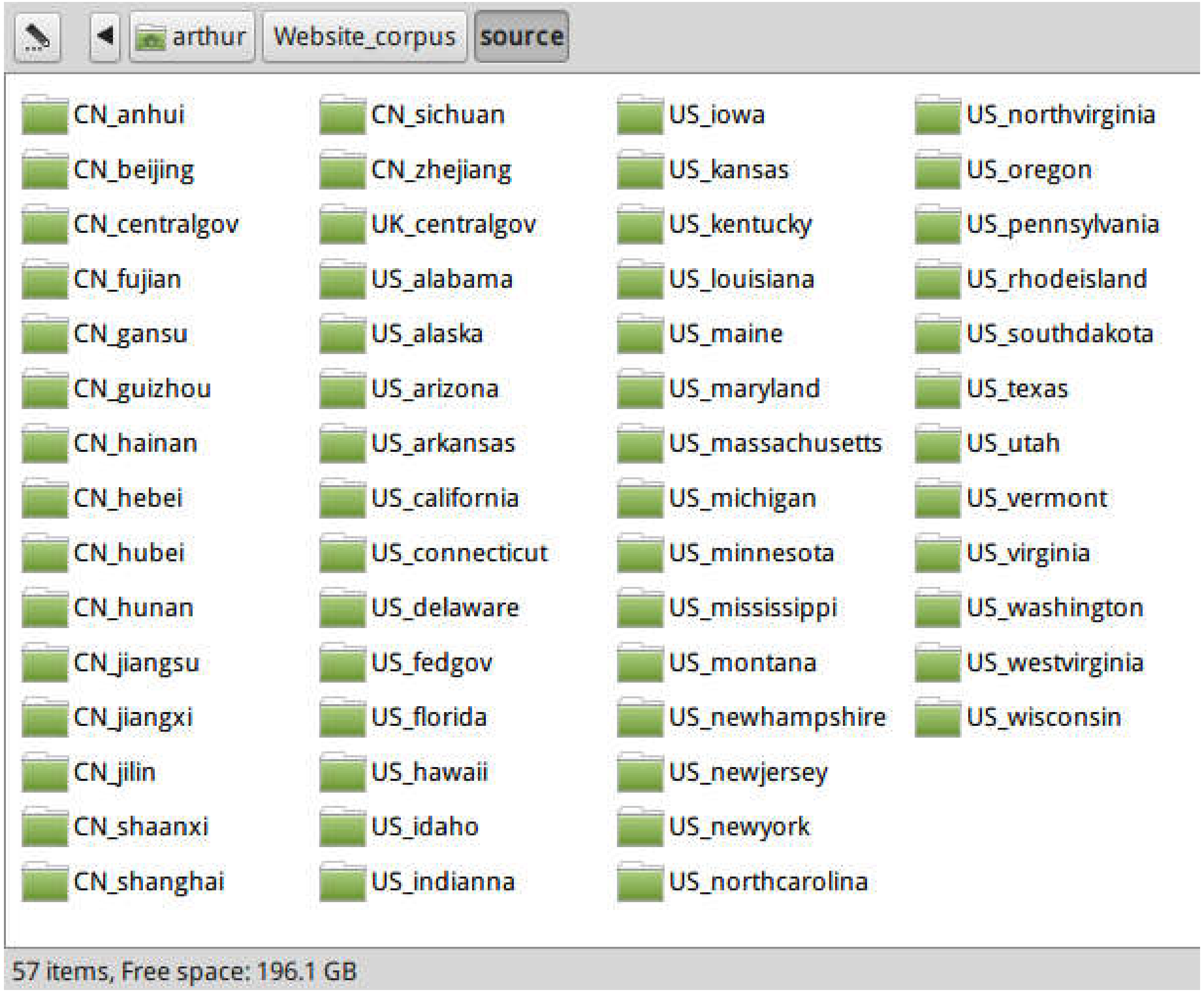
Corpus building

Step 1: downloading the websites

- Search the Internet for the URL of each websites in question;
- Download the websites using wget;
- Place the downloaded websites in a single folder called 'source';
- Rename the top folder name of each website, e.g. UK_centralgov.

Major website download tools (offline browsers):

- wget
- HTTrack
- SketchEngine WebBootCaT
- WordSmith Webgetter
- TelePort Pro



CN_anhui
CN_beijing
CN_centralgov
CN_fujian
CN_gansu
CN_guizhou
CN_hainan
CN_hebei
CN_hubei
CN_hunan
CN_jiangsu
CN_jiangxi
CN_jilin
CN_shaanxi
CN_shanghai
CN_sichuan
CN_zhejiang
UK_centralgov
US_alabama
US_alaska
US_arizona
US_arkansas
US_california
US_connecticut
US_delaware
US_fedgov
US_florida
US_hawaii
US_idaho
US_indianna
US_iowa
US_kansas
US_kentucky
US_louisiana
US_maine
US_maryland
US_massachusetts
US_michigan
US_minnesota
US_mississippi
US_montana
US_newhampshire
US_newjersey
US_newyork
US_northcarolina
US_northvirginia
US_oregon
US_pennsylvania
US_rhodeisland
US_southdakota
US_texas
US_utah
US_vermont
US_virginia
US_washington
US_westvirginia
US_wisconsin

57 items, Free space: 196.1 GB

Guestbooks and “printer friendly” directories were removed from the website folders before subsequent processing.

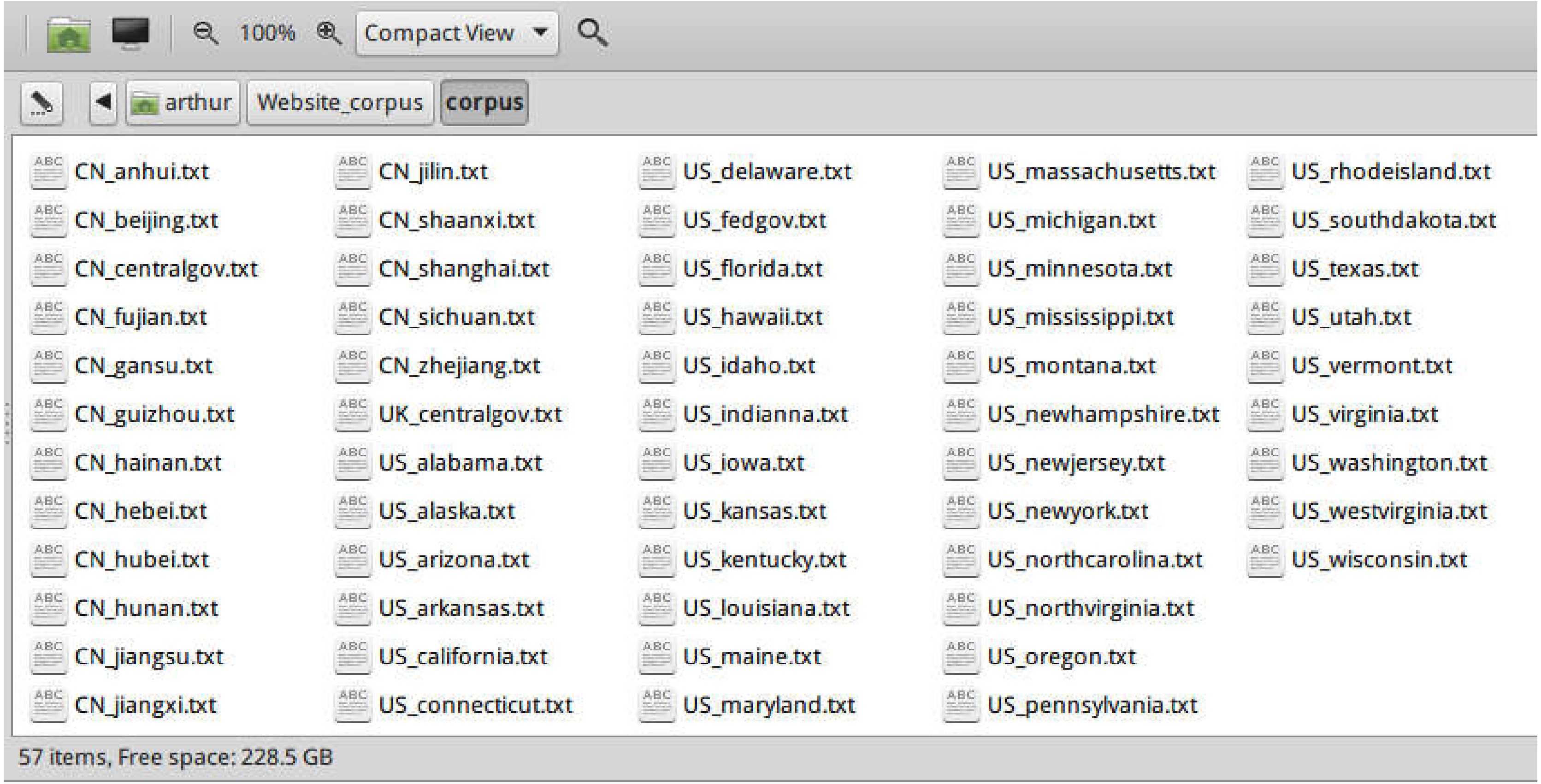
Corpus building

Step 2: extracting texts from html files

- Programming a tool;
- Use the tool to extract texts from html/xhtml/shtml files, so that all texts from a particular website are output to a single txt file e.g. 'UK_centralgov.txt';
- All txt files are written into a single folder called 'corpus'.

Text extraction

Remove preambles;
Remove scripts and css definitions;
Ignore list items within <selection>...</selection> blocks;
Render all headers, list items, options, independent phrases etc. as separate paragraphs;
Treat all text within one minimal tag pair as a single paragraph;
Ignore non-ASCII characters.



CN_anhui.txt
CN_beijing.txt
CN_centralgov.txt
CN_fujian.txt
CN_gansu.txt
CN_guizhou.txt
CN_hainan.txt
CN_hebei.txt
CN_hubei.txt
CN_hunan.txt
CN_jiangsu.txt
CN_jiangxi.txt
CN_jilin.txt
CN_shaanxi.txt
CN_shanghai.txt
CN_sichuan.txt
CN_zhejiang.txt
UK_centralgov.txt
US_alabama.txt
US_alaska.txt
US_arizona.txt
US_arkansas.txt
US_california.txt
US_connecticut.txt
US_delaware.txt
US_fedgov.txt
US_florida.txt
US_hawaii.txt
US_idaho.txt
US_indianna.txt
US_iowa.txt
US_kansas.txt
US_kentucky.txt
US_louisiana.txt
US_maine.txt
US_maryland.txt
US_massachusetts.txt
US_michigan.txt
US_minnesota.txt
US_mississippi.txt
US_montana.txt
US_newhampshire.txt
US_newjersey.txt
US_newyork.txt
US_northcarolina.txt
US_northvirginia.txt
US_oregon.txt
US_pennsylvania.txt
US_rhodeisland.txt
US_southdakota.txt
US_texas.txt
US_utah.txt
US_vermont.txt
US_virginia.txt
US_washington.txt
US_westvirginia.txt
US_wisconsin.txt

57 items, Free space: 228.5 GB

Excerpt from UK_centralgov.txt:

Get support from Jobcentre Plus to help you prepare for, find and stay in work, including:
training, guidance and work placement programmes
work experience, volunteering and job trialling schemes
help with starting your own business
help combining work with looking after children or caring responsibilities
extra help for specific problems
You may also be able to keep getting some benefits once you start working .
Support for disabled people
Speak to a Disability Employment Adviser (DEA) at your local Jobcentre Plus. They can help you find a job or gain new skills , and tell you about specific programmes to help you back into work.

Excerpt from US_southdakota.txt:

Board of Regents
Education Funding
Send a Postcard
Here is an opportunity to share just how beautiful and majestic South Dakota is with your friends and family. Send a virtual postcard and invite them to check out more of what the State of South Dakota has to offer by visiting sd.gov . First, choose any one of the postcards below, then complete the form indicating your contact information as well as the recipient's and finally include. First, choose any one of the postcards below, then complete the form indicating your contact information as well as the recipient's and finally include a brief and friendly message to go along with the postcard image. Best of all, it is free! No postage necessary.
Your Postcard Photo:Prairie Sunset

Excerpt from CN_hubei.txt:

Bao Yilin, a junior student from Wuhan Polytechnic University told the reporter that she finished the 7-day volunteering teaching in Cambodia not long ago. Learn More
Spring Airlines launches Wuhan-Tokyo service
Chinese budget carrier Spring Airlines on Saturday launched direct flights between Wuhan, capital of central China's Hubei Province, and Tokyo of Japan. Learn More
Visitors enjoy plum blossoms in Wuhan
More than 340 species of plum blossoms were planted in the East Lake plum blossom garden in Wuhan, central China's Hubei Province. Learn More
Chinas Taoist mountain opens airport
An airport near Mount Wudang, a destination known for its Taoist temples in central Chinas Hubei Province, started operation on Feb.5, 2016. Learn More

Size of the corpus

These were calculated case-insensitive:

	Types	**Tokens**	**TTR**
CN	126,115	21,692,518	0.0058
UK+US	213,695	60,199,969	0.0035
Total	339,810	**81,892,487**	

Keywords (obtained using COCA word list for reference)

1	china	2657343.423
2	s	2471491.706
3	hunan	2070108.103
4	shall	1540710.507
5	province	1485077.375
6	investment	996197.732
7	government	970200.671
8	cn	809228.383
9	chinese	775220.957
10	anhui	773066.711

1	classes	6371496.572
2	gov	2252689.336
3	services	2213016.158
4	s	2134123.288
5	state	1899258.608
6	department	1696316.792
7	resources	1505201.554
8	information	1438059.430
9	contact	1219153.098
10	utah	1203723.802

2 cases: *should*, *must*

	should		***must***	
	Freq	**/mln**	**freq**	**/mln**
CN	21,343	984	8,030	370
UKUS	19,711	327	33,705	560

should in CN subcorpus

ay Weihe River water quantity dispatching schemes should be formulated and released to low levels

al integrated infrastructure network. Besides, we should also make good plan for protecting ecologic

hority: within the time limit, effective measures should be taken to send the Certificates of

comprehensive profits of the tourism industry. We should constantly boost the transformation and upg

than the national average level. More efforts should be devoted to market supervision, social go

of mountains, waters, forests, lands and lakes should be carried out to fully promote green

speed and efficiency of economic growth. We should keep an eye on the orientation of

inspire consumption demands by all means. We should implement the linkage mechanism that works

should in UKUS subcorpus

tingling in fingers and toes. If you should experience any symptoms after water recreat

adds to your toolbar and our feeds should display properly in Chrome.

ling or completing Articles of Incorporation, you should first contact the IRS.

questions. If the person is legitimate, they should not hesitate to answer any questions you

only, that if the seventieth (70th) day should fall upon a Saturday, a Sunday, or

otice customarily forwarded to the depositor, and should include a copy of the dishonored instrument

creditor does not agree that the judgment should be satisfied/released, the judgment credito

mentary 1. Presence Of Both Parents. Both parents should be present at the time of the

must in CN subcorpus

bsidy policies, and enhance social assistance. We must not allow price increases to affect the

y for Admission Letter. c. Recommended candidates must finish the online application procedure at CS

. Chapter II Marriage Contract Article 5 Marriage must be based upon the complete willingness of

interests and public interest of society and must not take profit making as the object.

nt for fair competition. Procedures and processes must be simplified and time frames must be

and the changing society, and we artists must create new stuff to attract audiences." He

Industry, said the Belt and Road countries must not have any doubts concerning the need

Group, said that the Chinese liquor industry must move forward along with the culture of

must in UKUS subcorpus

Herzegovina - On or after November 20, 1995 (and must have begun on or before December 20, 1998). 1

wish to amend a nonresident return, you must file Form NJ-1040NR for the appropriate

that can be transferred without a waiver, must, nevertheless, be reported on a decedent's

eriod for transition of care. Prior authorization must be obtained when required. Beginning May 1, 2

then be available for your selection. Nominations must be registered on-line on or before 4:00

ol districts. School Year 2015-2016 All districts must submit Statement of Assurance (SOA) files thr

south of the designated bathing area. Fires must be at least 50 feet east of the

document added to this website. All vendors must watch the Division of Purchase and Property’

Speculation

The use of *should* is strikingly more frequent in the CN subcorpus; the use of *must* isn't very different from that in UKUS English, though both are possible translations of *yao* (要) or *bixu* (必须) in Chinese.

- The modal verb *yao* apparently reflects the influence of the former Soviet discourse, particularly the Russian word *nado*.
- *yao* in contemporary Chinese is translated more frequently into *should* than into *must* in English.
- *nado* → *yao/bixu* → should/must → should

Examples from PR China government report parallel corpus, 2000-2012

对于这些问题，我们要继续采取有力措施，切实加以解决。

We must continue to take effective measures to solve these problems.

进一步发挥货币政策的作用。

We should take better advantage of the role of monetary policy.

要切实改进金融服务。

Financial services should be substantially improved.

Further speculation

Should and *must* seem to be used interchangeably as translations of *yao* though this needs to be investigated using parallel corpora. (Unfortunately the corpus 'OPUS2 Chinese Simplified' on SketchEngine doesn't include “text type” or genre information.)

must and *should* are used very often in patterns like “we must/should ...”, which is typical of Chinese political discourse. Due to the hierarchical structure of the organization of the Chinese governmental system and the “unidirectionality” of government orders, this pattern is widely used in official regulatory documents.

Case: propaganda in CN subcorpus (272, 12.539/mln)

deal with areas such as ideology and propaganda, but the third one usually focuses on

the CPC Hainan Provincial Committee, Director of Propaganda Dept. of Hainan Committee of CCP and

Hunan Provincial CPC Committee & minister of the Propaganda Department of Hunan Provincial CPC Comm

co-sponsored by the Provincial Party Committee Propaganda Department, the Provincial Literary Fed

coordinate the liaison contact and the government propaganda with the news media; contact the Provin

coordinate the liaison contact and the government propaganda with the news media; contact the Provin

labor; 5. to formulate the plans on the propaganda of legal system and the popularization

rnment administrative departments of information, propaganda and communications shall, according to

th League I. To strengthen the propaganda and education of legality for the young

of a non-planned childbirth to run propaganda and education, and mobilize the parties

shall do a good job in the propaganda and explanation to the enterprises and

Case: propaganda in UKUS subcorpus (8, 0.1329/mln)

conforme anunciado. Se for feito erro em propaganda, obrigao da empresa fazer correes, e, a
caractersticas da garantia devem estar claros. No propaganda enganosa se o estoque do produto na
Compra Lembre-se de ler toda a propaganda atentamente. Se voc tiver dvidas a resp
coisa pela internet, de um catlogo, de propaganda impressa, por telefone ou solicitao pel
his role as defense counsel. Patriot-created propaganda included Revere's famous engraved print
regarding a political party and no political propaganda or campaign materials of any kind. Do
civil defense activities, and a Rumor and Propaganda Division whose task was to receive and
validity and thus undermine and minimize enemy propaganda efforts in Oregon. The end of the

Collocates (3–N–3) in CN subcorpus

MI:

1	2	2	0	10.39792	childbirth
2	2	2	0	9.24289	ideological
3	2	2	0	9.20527	propaganda
4	2	2	0	9.00733	admitted
5	2	2	0	8.31832	recognition
6	2	2	0	7.70965	experiences
7	2	2	0	7.21325	feature
8	2	2	0	6.68356	television
9	2	2	0	6.55127	authorized
10	5	5	0	6.51007	minister
11	2	2	0	6.46619	household
12	2	2	0	6.33346	publicity
13	5	5	0	6.06753	carry

T-score:

(None of the “collocates” scored higher than 4.)

Collocates (3-N-3) in UKUS subcorpus

Task failed successfully.

Trigrams (all data treated as lowercase)

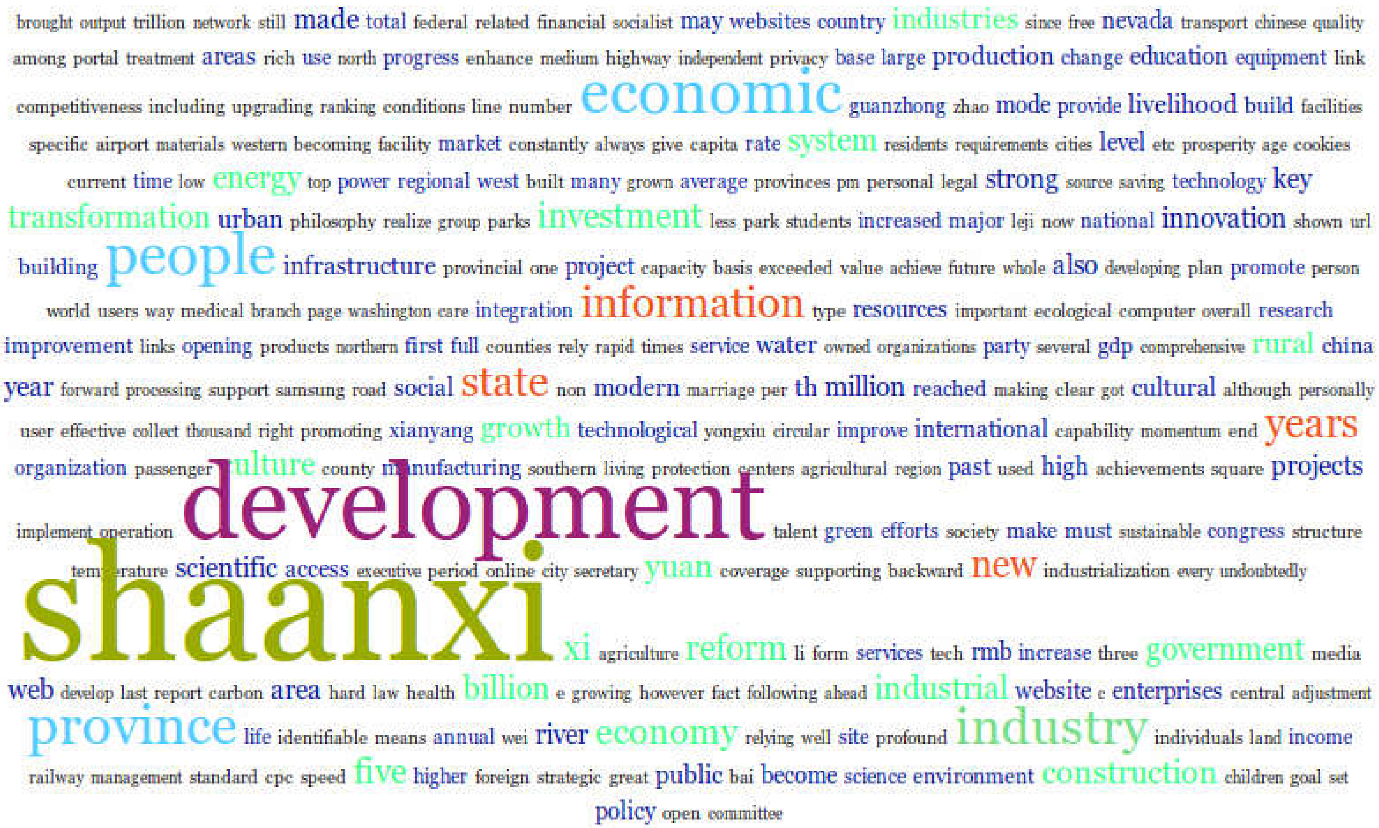
CN subcorpus:

1	29677	all rights reserved
2	23807	the people s
3	19689	people s government
4	14589	s government of
5	13922	province all rights
6	12027	the peoples government
7	11926	peoples government of
8	11224	of jiangxi province
9	11115	the state council
10	11058	www gov cn
11	10663	government of jiangxi
12	10625	according to the
13	10406	in accordance with
14	9908	as well as
15	9038	the work of

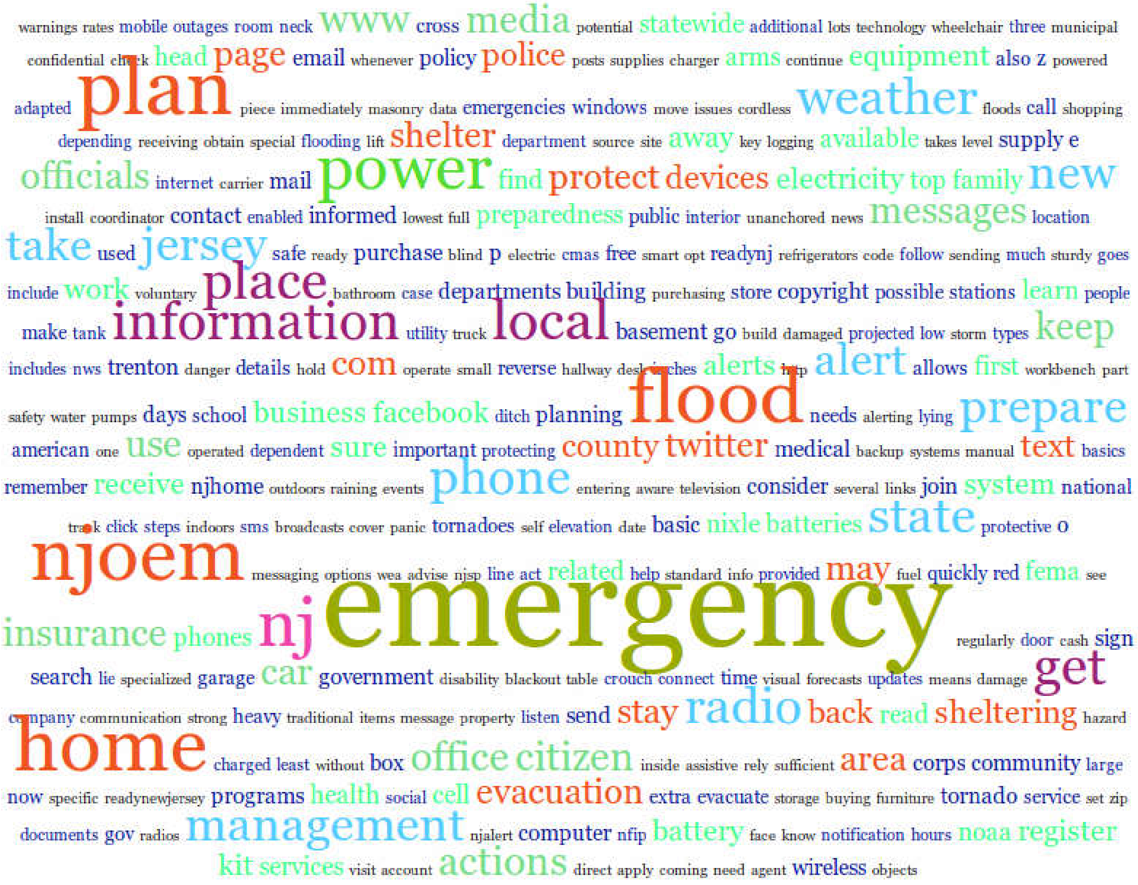
UK/US subcorpus:

1	58862	department of commerce
2	56565	continuing education provider
3	54181	total votes pct
4	46880	skip to content
5	44625	frequently asked questions
6	41076	office of the
7	33062	secretary of state
8	30322	provided by the
9	29376	a a a
10	29020	utah department of
11	28814	size a a
12	28810	font size a
13	28575	utah division of
14	28546	a to z
15	28467	links and resources

“Word cloud” of a typical CN corpus text



“Word cloud” of a typical US corpus text



“Wordprofile” (obtained using COCA 25-level word lists)

Word profile of CN

LEVEL	TOKEN	TOKEN%	GROUP	GROUP%
1	12868201	58.514	1000	7.171
2	3250655	14.781	1000	7.171
3	2482613	11.289	1000	7.171
4	484372	2.203	998	7.156
5	219553	0.998	992	7.113
6	126504	0.575	969	6.948
7	71890	0.327	947	6.790
8	101142	0.460	886	6.353
9	34576	0.157	837	6.002
10	34254	0.156	738	5.292
11	21411	0.097	670	4.804
12	17922	0.081	608	4.360
13	9766	0.044	520	3.729
14	15367	0.070	426	3.055
15	4867	0.022	397	2.847
16	6669	0.030	326	2.338
17	2890	0.013	287	2.058
18	2878	0.013	233	1.671
19	2150	0.010	247	1.771
20	2112	0.010	194	1.391
21	1757	0.008	160	1.147
22	1809	0.008	157	1.126
23	1635	0.007	132	0.947
24	977	0.004	111	0.796
25	1148	0.005	111	0.796

Num. of tokens not found in the lists: 2224404 (10.11%)

Word profile of UK|US

LEVEL	TOKEN	TOKEN%	GROUP	GROUP%
1	32253743	52.751	1000	6.584
2	9868107	16.139	1000	6.584
3	6560623	10.730	1000	6.584
4	1914639	3.131	999	6.577
5	592217	0.969	995	6.551
6	421422	0.689	974	6.413
7	382811	0.626	940	6.189
8	172667	0.282	908	5.978
9	149046	0.244	850	5.596
10	120449	0.197	782	5.148
11	60913	0.100	727	4.786
12	37746	0.062	651	4.286
13	25566	0.042	554	3.647
14	63014	0.103	539	3.549
15	16169	0.026	455	2.996
16	20653	0.034	402	2.647
17	8433	0.014	382	2.515
18	33874	0.055	361	2.377
19	4489	0.007	316	2.080
20	5867	0.010	283	1.863
21	5051	0.008	254	1.672
22	5660	0.009	245	1.613
23	3433	0.006	216	1.422
24	8429	0.014	205	1.350
25	3531	0.006	151	0.994

Num. of tokens not found in the lists: 8404875 (13.75%)

Todo

Download the websites again (done)

- To make sure the corpus includes all the websites of interest

Remove the 'noise' in corpus (solved with jusText tool by Jan Pomikálek)

Do re-sampling

- Even with the noise parts removed, the corpus will still be very large. For the purpose of manual annotation, re-sampling is necessary beforehand

Annotation

- Biber's (1988) model: multidimension analysis