CLARIN-PLUS workshop “Creation and Use of Social Media Resources”
Kaunas Hotel, Lithuania, 18 -19 May 2017

On the challenges of cross-national comparative research of NLP

Jiyoung Ydun Kim
PhD student
Anja Bechmann
Associate Professor
AU DATALAB
School of Communication and Culture
Aarhus University, Denmark

Background

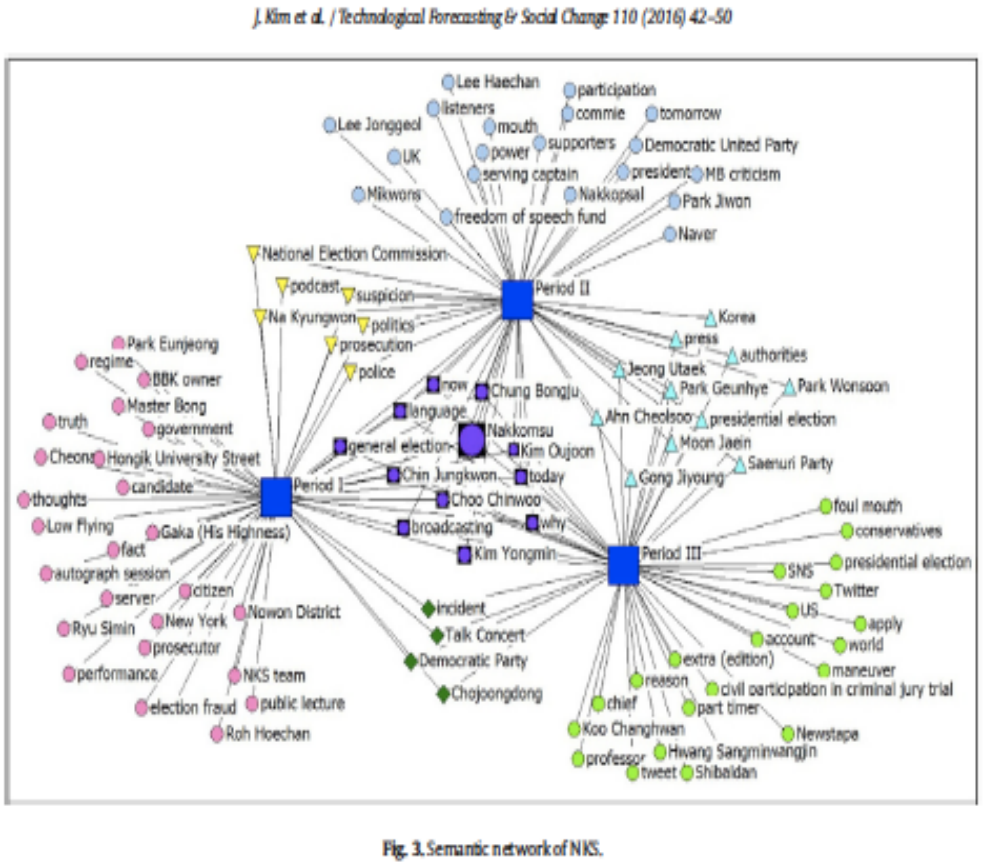
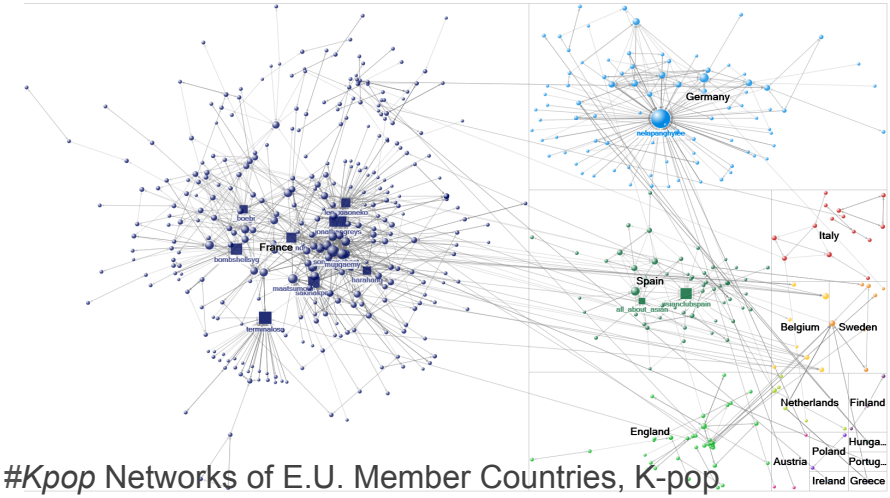


Fig. 3. Semantic network of NKS.



#Kpop Networks of E.U. Member Countries, K-pop communication network closely formed five degree of communication, with an average geodesic distance of 5.053.

	ID	BC	Ego network 1.5	Ego network 3.0	Ego network 4.5	Country	Type of Users
1	seoulfm	2879 2733				Philippines	Radio channel
2	soundrhythm	4917 407				Indonesia	Music Production
3	K-pop_lov	4031 056				Japan	K-pop Fan
4	kstarnews	3096 452				Japan	K-pop Fan
5	kannnnnnnnnnm	2995 266				Japan	K-pop Fan
6	sundaymandarin	2809 855				Indonesia	Radio channel
7	looxney	2709 486				France	K-pop Fan
8	intan_rha	2558 390				Indonesia	K-pop Fan
9	buchaan0222	1759 812				Japan	K-pop Fan
10	leeteuk_amour	1749 686				Japan	K-pop Fan

Political communication, Digital Journalism, Information network pattern, Digital divide. Cultural analytics
Social network analysis

Using Social Netowrk Analysis Programs
#NodeXL #Ucinet #Automap #KrKwic ..

Research Framwork

RQ		Concepts & Operational definition	Method
Group Characteristic -open* -closed* -secret*	- Group profile	- Group age& size (unique group member who post, like, comment, share) - Level of Interaction: bases on different types of post (photo, link, status, video and others) - The number of like per post - The number of comment per post - Time taken to receive like - Time taken to receive comment - Finding Facebook group Dunbar's number*: - Group size (network size) and communication interaction	Descriptive Statistics -Min. / Max. -Average -Median -Mode -Std. Deviation -Variation -Skewness -Kutosis Analysis of Variance (ANOVA)
Group network & Group Topics	- Group Structure - Semantic structure	1) Node (vertex): user / word 2) Link (edge): the relationship: liking, commenting, sharing/ co-occurrence - Co-commenter's network - Co-liker's network - Co-sharing network - Post semantic network - Comment semantic network - URL network 3) Network centrality index - Density: the ratio of the actual link in a network to the number of maximum possible edges	Social Network analysis Semantic network analysis Web Impact analysis
Group communication classification and prediction		After finding important variables based on the basic and network characteristic of group communication Classification :Bayesian classifier, logistic regression, KNN Classifier, Support Vector	Machine learning
Cross-country comparison		Korean network classification vs. Danish network classification	Cultural analytics

"Do we have the same Dunbar' numbers online?
"What is the relationship between the topics and the communication network size? on group communication?

1 Gender Social Capital Inequality on Facebook Groups:
a cross-country comparative study between Denmark and South Korea

Jiyoung Ydun Kim & Anja Bechmann

2 **What we use Facebook groups for**

Anja Bechmann, Jiyoung Ydun Kim & Anders Søgaard

*Preprocessing steps for the Topic modeling

social cohesion

- Experience radical individualisation and challenges of collective voices on international level (eg. Brexit, climate changes)

- Common values are eroding (if ever existed) – on Facebook international different values become visible (eg. Censorship) – how are different cultures using Facebook – similar or different?

- Macrostructures are potentially changing, but at the same time group structures/relational selves blooms

(REF: Bauman, Simmel, Habermas)

FB groups

- 900 million use WhatsApp, 400 million use Instagram, 700 million use Messenger and 700 million use Facebook Groups.
- According to Mark Zuckerberg pointed out that Facebook groups is one of Facebook's core products as it provides an option for sharing with either closer users or larger communities. Furthermore, Facebook group service is one of the top three visions of the company for the future, together with Messenger and Instagram. .
- Three types of groups according to visibility/privacy settings on FB: **open, closed, secret**

Despite the overwhelming amount of users of Facebook Groups, we know very little if anything on the variety of topics within these groups.

What are people actually using the groups for, what kind of topics are they discussing, communicating, and connecting around?

How does this differ according to group privacy settings, gender and nationality?

What can we potentially learn from Facebook groups on the topic of social cohesion and the theory of social groups

- Facebook groups have been the focus of several existing studies, but they have focused on content in particular public and political groups (e.g. Fernandes et al., 2010; Marichal, 2013).
- Others focus on sense making in personal groups through the use of surveys (e.g. Namsu, Kee & Valenzuela, 2013) or ethnographic studies (e.g. Miller, 2011).
- The study of content patterns in personal groups in this paper will add to this knowledge within the field of internet research.

Korea and Denmark were chosen since both countries have a high Internet and Facebook penetration, but differ in gender equality (GDI, GEM, GII Index) and different cultural contexts

This allows for a comparative level of potential posting frequency, but on the other hand also ability for the study to detect and discuss expected cultural differences in topics that is not possible in most similar case study designs.

	Denmark	Korea
Population	5,711,837	50,704,971
Internet users	5,479,054(96%)	45,314,444(89.4%)
Own smartphone	83%	84.3%
Facebook Users	3,700,000	17,000,000
Social media users	76,7% of the DK population 12+ years has a social media profile	73.1% of Korean internet user use social network service;
	Facebook is the most used social media in Denmark with 97% of all social media users using Facebook_Other social media platforms ranked according to most frequent use is Google+ (41%), Snapchat (31%), Instagram (31%), Twitter (16%) and Pinterest (9%) (Ministry of Culture, 2015).	Majority of social profile-based social network service is KaKao stories (45.7%), Facebook (30%), Twitter (10.8%), Naver band (7.2%) . While male prefer the open SNS such as Facebook, Twitter, women prefer to use relatively closed type of visual SNS such as Kakao story, Instagram

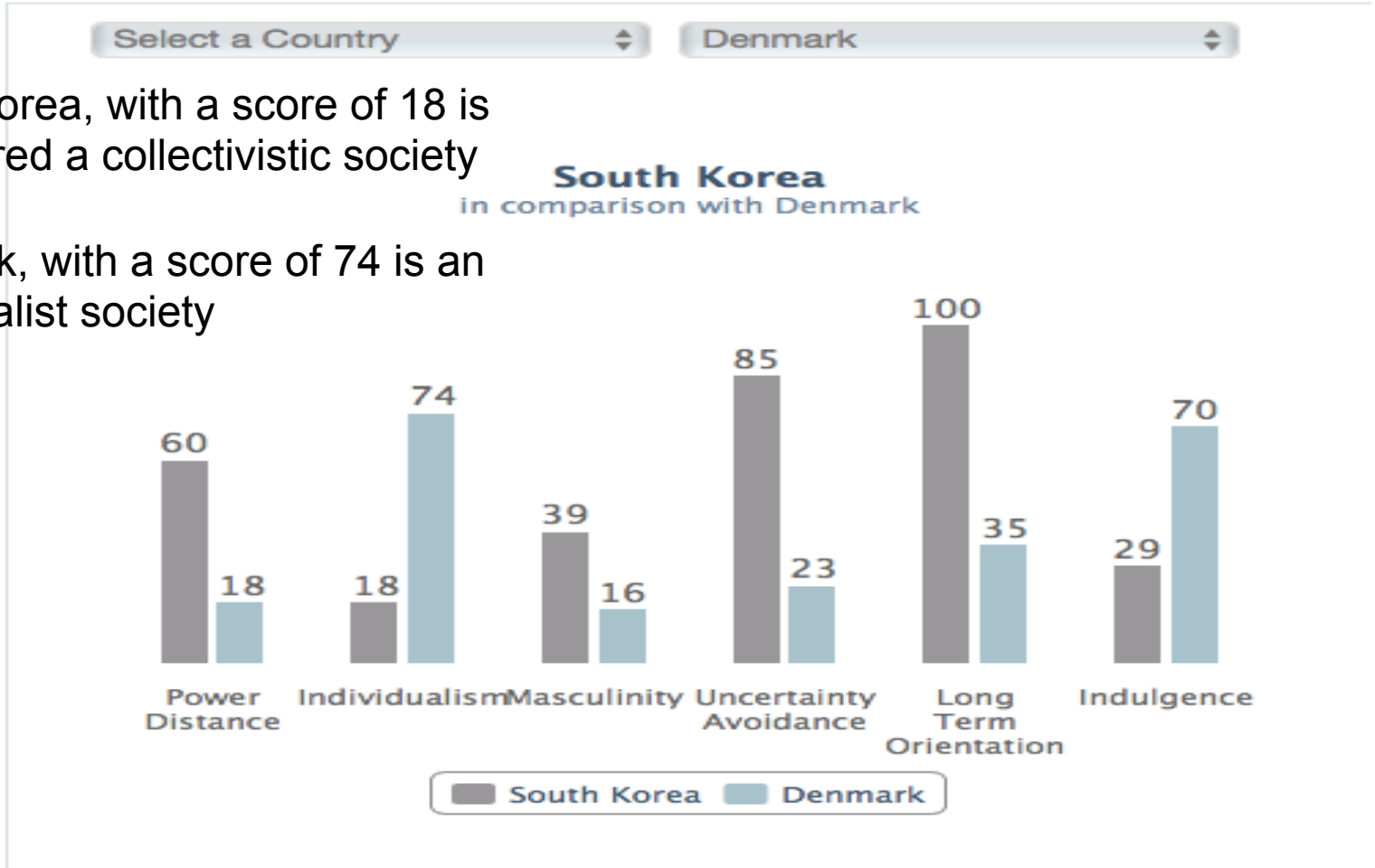
Note. Adapted from Internet World Survey(2016)

Cultural differences (Hofstede)

https://geert-hofstede.com

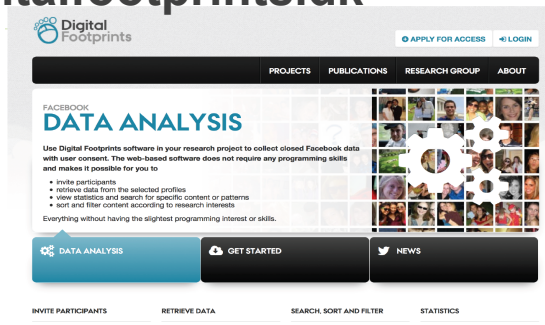
South Korea, with a score of 18 is considered a collectivistic society

Denmark, with a score of 74 is an Individualist society



Data Collection

http://www.digitalfootprints.dk



- **Digital Footprints Program**

•Digital Footprints (Denmark, Arhus University)
- API Data collection (see Bechmann & Bahlstrup, 2015)

•The Digital Footprints software (www.digitalfootprints.dk) is a data extraction and analytics software that allows researchers to extract user data from Facebook and Instagram data sources, public streams, as well as private data with user consent. Digital Footprints supports recruiting participants for research projects and then ask for permission to use their content. In this way researchers are able to control the demographics and to secure participation (Bechmann & Vahlstrup, 2015).

•Data period: 2014 May – 2015 April (until FB change the API privacy policy)

Data Collection

In regard to research ethics and legal issues,

This research is approved by the Social Behavioural Research Institutional Review Board (SBR, IRB approval number: 7002016-A-2015-002), for the Korean data, and by the Danish Data Protection Agency, for the Danish data.

Meta Data:
1000 Danes and 1121 Koreans to mirror the estimated national Facebook population of the two countries. The recruitment took place through internet panels and the strata has primarily been gender, age and educational level. .

we have collected the communication (posts and comments) within the secret, closed and open groups that our samples are members

- **Screenshot of Authorize Facebook App**

Data Collection

- **Collected Data DB**

	A	B	C	D	E	F	G	H	I	J
1	Id	From	From Id	Type	Post Type	Likes Count	Shares Count	Privacy	Created	Updated
2	22841474	U00-00-1152	17696079	POST	status	0	0		00:59.0	00:59.0
3	22841473	U00-00-1152	17696079	POST	status	0	0		01:34.0	01:34.0
4	22841472	U00-00-1152	17696079	POST	link	0	0		27:14.0	27:14.0
5	22841471	U00-00-1152	17696079	POST	link	2	0	EVERYONE	46:13.0	45:01.0
6	22841470	U00-00-1152	17696079	POST	link	5	0	EVERYONE	47:47.0	47:47.0
7	22841469	U00-00-1152	17696079	POST	link	0	0		44:26.0	44:26.0
8	22841468	U00-00-1152	17696079	POST	status	0	0		24:38.0	53:34.0
9	22841424	U00-00-0506	113519149	COMMENT	null	2	null	null	59:58.0	null
10	22841339	U00-00-0276	17695203	POST	video	2	2	ALL_FRIENDS	35:52.0	35:52.0
11	22841467	U00-00-1152	17696079	POST	status	0	0		46:55.0	46:55.0
12	22841157	U00-00-1152	17696079	POST	link	0	0	CUSTOM	43:32.0	43:32.0
13	22841156	U00-00-1152	17696079	POST	link	0	0	CUSTOM	49:43.0	49:43.0
14	22841464	U00-00-1152	17696079	POST	status	2	0		15:42.0	15:42.0
15	22841463	U00-00-1152	17696079	POST	status	30	0	EVERYONE	18:20.0	45:03.0
16	22841461	U00-00-1152	17696079	POST	link	0	0		26:53.0	26:53.0
17	22841460	U00-00-1152	17696079	POST	link	1	0	EVERYONE	27:36.0	27:36.0
18	22841459	U00-00-1152	17696079	POST	status	0	0		02:15.0	02:15.0
19	22841457	U00-00-1152	17696079	POST	photo	0	0	EVERYONE	38:10.0	38:10.0
20	22841456	U00-00-1152	17696079	POST	status	0	0		33:11.0	33:11.0

Anonymization of user information

Descriptive result

	Danish Data				Korean Data			
Privacy settings	Open Group	Closed Group	Secret Group	Total_group	Open Group	Closed Group	Secret Group	Total_group
Number of groups (%)	5,730 (39)	6,669 (46)	2,209 (15)	14,608 (100)	3,029 (49)	2,087 (34)	1,059 (17)	6,175 (100)
Total number of posts (%)	1,806,306 (21)	5,665,811 (67)	938,235 (11)	8,410,649 (100)	1,139,369 (51)	843,520 (37)	268,017 (12)	2,250,906 (100)
Average posts	315.2	849.6	424.7	575.7	376.6	404.2	253.3	364.7
Mean posts	96	72	26	69	85	18	16	37
Total number of comments (%)	4,520,810 (12)	26,974,589 (74)	4,852,620 (13)	36,348,019 (100)	1,475,199 (17)	5,509,287 (66)	1,419,522 (17)	8,404,312 (100)
Average Comments per group	788.9	4044.8	2196.7	2488.2	487.0	2639.8	1340.7	1361.0
Mean Comments	144	176	71	137	42	32	25	35

Danes are members of twice as many groups as in Korea
Danes use more closed group whereas S. Korea use open groups
The average and mean posts and comments are higher in DK than in S. Korea indicatin
a more frequent use

Friends_Descriptive result based on gender and education

Table. Descriptive result based on gender and education.

Variable	Variable	N	Min.	Max.	Mean	S.D.	Median	Skewness	Kurtosis
Gender	Female	555	0	4878	272.78	383.21	208	7.8	80.1
	Male	566	0	4832	416.92	651.50	266.5	4.5	23.5
Education	Short	601	0	2887	247.6	216.8	210	4.2	41.9
	Medium	380	0	4757	324.6	473.2	227.5	6.0	47.3
	Long	140	15	4878	820.2	1125.5	386.5	2.3	4.9
Sum	Sum	1121	0	4878	345.56	540.33	230	5.5	37.0

Korean_result

- The obtained results showed that there are statistically significant differences of the gender and between the two countries.
- In Korea, men had in average 1.5 times more friends than women, 416.9 friends compared to 272.8, respectively, whereas women has more friends that men, 260.9 friends compared to 197.0 in Denmark

Preliminary results

- There was also a noticeable difference in the group membership gaps between genders from both countries. The Korean men were members of 7.1 groups in average while women were members of only 2 groups in average.
- In Denmark, with women subscribing more groups than men (20.6 for women and 13.3 for men).
- Denmark is closed (8.3), open (7.0) and secret (2.3); whereas in Korea is open (5.0), closed (2.3), and secret (1.3).

Preprocessing steps for the Topic modeling

Our preprocessing steps are the same for the two languages, ignoring the many linguistic differences between the two languages**. We follow the same steps not to introduce systematic biases that could complicate our comparative analysis.**

- Remove
- (i) **words with less than three characters**
- (ii) words that occur in the **NLTK stop word lists**, including, in the case of Danish, the most frequent words in the Corpus 2000 corpus of Danish

- tokenization for Korean corpus :Polyglot out of Google vs. other options

- (iii) words that do *not* occur in the Danish-English and Korean-English Wiktionary bilingual dictionaries.
- The third filter (iii) limits our vocabulary significantly and introduces a potential bias toward topics with higher coverage, but on the other hand, it also removes most, if not all, noise from the signal.
- Latent Dirichlet Allocation (LDA) : interested in what is being said and possible discursive/topical shifts

Questions.

Any attempt to standardize NLP methodology among different languages?

How can we measure or test reliability and credibility on multilingual analysis?

CLARIN-PLUS workshop “Creation and Use of Social Media Resources”
Kaunas Hotel, Lithuania, 18 -19 May 2017

On the challenges of cross-national comparative research of NLP

Jiyoung Ydun Kim and Anja Bechmann

AU DATALAB
School of Communication and Culture
Aarhus University, Denmark