Digital Disruption: A Corpus-based Analysis of Keywords in the Bangkok Post Newspaper (January 2017–February 2018)

*Jonathan Rante Carreon[1]
Banjob Piyamat[2]
Faculty of Liberal Arts, Huachiew Chalermprakiet University[1,2]
Samutprakarn, Thailand
Corresponding Author's Email: carreonjrc@gmail.com

Abstract

The ultimate goal of this research study is to investigate the information relating to the phenomenon of digital disruption that is presented to the public by the media. Employing mixed method of analysis, 81 news articles that were reported from January 2017 to February 2018 in Thailand's leading English newspaper, *The Bangkok Post*, composed of 101,895 words, were examined for keywords. Using the free corpus tool *AntConc*, the investigation was done by comparing words, with a minimum absolute frequency of 100, against their frequencies in the British National Corpus (BNC) using log-likelihood (see Rayson & Garside, 2000 for details of log-likelihood uses). Any words with a log-likelihood (LL) value greater than 100 were considered keywords. The resulting keywords were iteratively thematized by each of the researcher, and the degree of inter-rater agreement for accuracy and reliability in categorization is expressed as Cohen's kappa value. The analysis yielded 34 keywords composed of words relating to digital technology and its applications (N=11; 32.35%), words relating to business and monetary issues (N=10; 29.41%), words relating to digital technology potential stakeholder (N=5; 14.71%), words relating to digital technology impact (N=3; 8.82%), words relating to location and time (N=3; 8.82%), and words relating to informational dimension of language (N=2; 5.82%), with the analysts categorization having a substantial level of inter-rater agreement (Cohen's kappa= 0.785).

Keywords: corpus-based analysis, digital disruption, digital technology, iterative thematic analysis, keywords

1. Introduction

The Digital Era has introduced different phenomena such as globalization, glocalization, digitalization and many others. One phenomenon that has a wide impact in today's society is digital disruption. Generally speaking, digital disruption may is a turning point in the lives of both the new and older generations, termed by Prensky (2001) as digital natives and digital immigrants, respectively. To say the least, on the one hand, young people, especially children, implicitly understand digital disruption in a way that the rest of us can only watch and admire since they were born in this digital era, where their immediate digitalized environment work as an instant teacher to them. In other words, they born and raised on digital technology in their homes, schools, and toys. On the other hand, digital immigrants, like all immigrants in its real sense, while may adapt to their new environment, "they always retain, to some degree, their 'accent,' that is, their foot in the past. … [w]hich can be seen in such things as turning to the Internet for information second rather than first, or in reading the manual for a program rather than assuming that the program itself will teach us to use it" (Prensky, 2001, p.2).

In the world of business and computing, the phenomenon of digital disruption may be used for enhancing competitiveness. For instance, there is the wide accessibility of tremendous amounts of information that are never seen before, ranging from proprietary data to new public sources of open data that may be useful in conducting and managing

businesses. Hirt and Willmott (2014) argued that analytical and processing capabilities have made leaps with algorithms scattering intelligence across digital networks not to mention the smart mobile devices, and the accessibility of information and computing power to users around the world. Despite these advantages they warned that "digitization is rewriting the rules of competition, with incumbent companies most at risk of being left behind" (Hirt & Hilmott, 2014, p.1). In other words, while digital disruption is beneficial to some sectors of the society, its aftermath is also quite common and needs to be equally made known to all concerned individuals.

For instance, in the field of business and economics, Bughin (2017) found that digitization has a significant negative impact on the profits of the traditional business operators through two loop effects: (1) digital entrants competing with traditional business operators through disruptive practices, and traditional business operators responding to disruption and creating more intense competition with each other (see also Gilbert (2015) on the impact of digital disruption on e-book sales of Amazon).

In education, McCoy (2016) examined classroom learning distractions caused by the use of digital devices for non-class purposes and found that respondents spent an average of 20.9% of class time using a digital device for non-class purposes.

In newspaper industries, Weber and Monge (2017) investigated the inclusion of hyperlinking as part of the transformation process from print-based organizations to multimedia information providers during an 11-year period from 1997 to 2007, and found that traditional newspapers which ignored hyperlinking practices had an increased likelihood of failure in the long run.

It is quite evident that the number of research studies on digital disruption has been an increasing, but these studies either focused only on the benefits or the negative impacts. The data investigated in most of these research enterprises were either taken from the company or from the result of interviewing concerned individuals. The authors argues that due the advent of advanced digital and computer technology, most readers may have switched already to reading online news items, so the data for this investigation is taken from a newspaper website.

2. Research Question and Conceptual Framework

The lone research question is "What information on digital disruption is presented to the readers?" To answer this research question, a corpus-based analysis examining "keyness" in text (Scott & Tribble, 2006) is employed. Keyness is "a quality that words may have in a given text or set of texts, suggesting that they are important and they reflect what the text is really about" (Scott & Tribble, 2006, pp. 55-56). This concept of keyness is the underlying notion behind the analysis of keywords (Scott, 1997).

Keywords are "words which appear in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size" (Baker, Hardie & McEnery, 2006, pp. 97-98). This means that keyword analysis is a comparison of one corpus against another, generally a comparative corpus (also known as a benchmark corpus; e.g. British National Corpus or BNC), in order to determine which words "occur statistically more frequently" in either of the corpora (Baker, 2006, p. 125). This takes into consideration relations between frequency and typicality among words (keywords) between two corpora (Stubbs, 2001). Thus, keywords may keywords highlight the important content of a text such as priorities (Carreon & Watson Todd, 2013).

Given the rapid turnover of information on digital technology and the inevitable impacts of the phenomenon of digital disruption to the local and international businesses practices (Yu & Hang, 2010), research studies that examine keywords may shed some light on specific information about that phenomenon, which may have implications in local and

international sectors such business, education, and even in daily living. In this research, such implicative information on digital disruption is examined from the news articles.

3. Related Research

Originally coined as disruptive technologies then later called disruptive technologies (Christensen, 1997), digital disruption refers to innovation that creates a new market and a network of valuing things, with this creation eventually disrupting an existing market and value network, and displacing established leading firms in the market, including their products, and sometimes their alliances (Christensen, 2013; see also Bradley & O'Toole, 2016). Downes and Mui (1998) posited that the process of digital disruption involves the "transformation of information into digital form, where it can be manipulated by computers and transmitted by networks" (p.6).

While it is known as the most influential business idea of the early 21st century, research studies on digital disruption are concentrated in the Western countries, examining mostly national electronic health records and the digital disruption of moral orders (e.g. Garrety, McLoughlin, Wilson, Zelle, & Martin, 2014); technological innovation and global film circulation (e.g. Iordanova, 2012; Iordanova & Cunningham, 2012); how digital disruption is redefining industries (e.g. James, 2013; Joseph et al., 2015; Jahangir & Walter, 2015); and digital disruption-connectivity and the diffusion of power (e.g. Eric & Cohen, 2010). The few research studies on digital disruption in Asia focused on the impact on ICT production (e.g. Wong, 2002); impact on running businesses (e.g. Wikstrom & De Fillippi, 2016; Tornjanski, Marinkovic & Savoiu, 2015); and perceptions and responses of affected individuals and industries to digital disruption (e.g. Karpathiou, 2016; Marous, 2013; Hicks, 2016).

In Thailand, the few research on digital disruption were mainly on promoting security and cultural awareness (Fung, Khera, Depickere, Tantatsanawong, & Boonbrahm, 2008; Hongladarom, 2016). Thus, it is interesting to know what information about digital disruption is presented to the end users of digital technology. Most information about digital technology and its applications are emanating from the industries and manufacturers. This research study examines digital disruption as presented in newspapers since newspapers are the most common type of publications read by Thais (Statista, 2018). In particular, the news articles on *Bangkok Post Online Newspaper* that are related to digital disruption are investigated for keywords.

4. Methodology

3.1 Data

Eighty one news articles composed of 101,895 words that were reported from January 2017 to February 2018 relating to digital disruption were downloaded from the *Bangkok Post Online Newspaper* and examined for the information presented during this period. The *Bangkok Post Online Newspaper* was chosen mainly because it has the highest number readership (Thongtep & Pratruangkrai, 2016). Moreover, online newspapers are one of the fastest sources of information available to public.

3.2 Analysis

Linguistic keywords reflect the content of a particular text (Scott, 1997, 2000) through their high frequency. In this study, linguistic keywords were identified by, first, conducting a word frequency count on the texts of the websites to find the absolute frequencies of all words. While absolute frequencies can provide some useful information about the concerns of a text,

in many cases the words with the highest absolute frequencies will be similar across different texts simply because these words are most commonly used in English (such as articles and prepositions). Therefore, relative frequencies of words compared to a benchmark of general English use are more insightful, and it is the words with high relative frequency that are considered keywords. In this study, the frequencies of all words (including British and American spelling variants together) in the data with a minimum absolute frequency of 100 were compared against their frequencies in the British National Corpus (BNC) using log-likelihood (see Rayson & Garside, 2000 for details of log-likelihood uses). Any words with a log-likelihood of greater than 100 were considered keywords. These 34 keywords within their local co-text were then categorized using an iterative process of identifying themes. Six themes were identified. These are:

- Words relating to digital technology and its applications
- Words relating to business activities and monetary issues
- Words relating to digital technology potential stakeholder
- Words relating to digital technology impact
- Words relating to location and time
- Words relating to informational dimension of language (Biber, Conrad & Reppen, 1998)

The first five of these categories shed light on the purposes of this study: words relating to business and monetary issues, words relating to digital technology and its applications, words relating to digital technology impact, words relating to place and time, and words relating to potential stakeholder of digital technology and its applications are all the kinds digital information that were chosen to be presented to the public through the online version of the newspaper. Words relating to informational dimension of language shed some light on the perspectives of underlying authors presented in the news as well as potential future directions of digital disruption in Thailand.

The two researchers then categorized the keywords into these six themes independently, and the categorizations were compared for reliability using Cohen's kappa. The Cohen's kappa is a statistical coefficient that represents the degree of accuracy and reliability in a statistical classification and measures inter-rater agreement by classifying items into mutually exclusive categories.

The following guidelines suggested by Landis & Koch (1977) can be used to interpret the results:

- 0.01 – 0.20 slight agreement
- 0.21 – 0.40 fair agreement
- 0.41 – 0.60 moderate agreement
- 0.61 – 0.80 substantial agreement
- 0.81 – 1.00 almost perfect or perfect agreement

4. Results and Discussion

4.1 Absolute frequency analysis

Table 1 Words with the highest absolute frequencies

Words	Frequency (f)	Words	Frequency (f)
the	3868	a	1230
to	2557	for	979
and	2285	is	836
of	2084	as	615
in	1556		

Some initial implications can be drawn from Table 1 above. The high frequency of the articles *the* (f=3868) and *a* (f=1230) reflects the presence of many long running sentences in the data where these articles were used as markers of specific and non-specific references, respectively (Quirk, Greenbaum, Leech & Svartvik, 1995; see also Nickalls, 2011). The frequent use of the prepositions such as *to* (f=2557), *of* (f=2084), *in* (f=1556) and *for* (f=979) function to link nouns, pronouns, or noun phrases to some other parts of the sentences. For instance, in the corpus, *to* is either used as preposition to express direction or motion or direction toward something or as a marker of an infinitive verb (e.g. *to* Asia, *to* drive); *in* is used to indicate either unspecific times during a day, month, season, year or used to indicate a location or place (e.g. *in* decade, *in* a country); and *for* is used to tell about the use of something, a reason or purpose (e.g. *for* a chance, *for* cryptocurrency).

The high frequency of the conjunction *and* (f=2285) in the corpus shows common use of conjoint words (e.g. brand *and* a company, airline *and* sizeable medical tourism sector), providing a series of examples, and as coordinating conjunction in compound sentences (e.g. shoes, bags *and* accessories, use the platform *as* a discount catalogue). The high frequency of the auxiliary verb *is* (f=836) indicates that the information presented in the news is current or fresh (e.g.*is* a big need, *is* the fastest growing). The high frequency of *as* (f= 615) is either used as a causal conjunction (e.g. *as* a critical quality, *as* a combination) or to show function (e.g. digital transformation *as* a corporate strategy, register it *as* a company, *as* a business-to-business (B2B). However, it is difficult to reach stronger conclusions as these words reflect general language use and some specific uses of the corpus. Thus, there is a need to examine the relative frequencies words by comparing the absolute frequencies against frequencies in the BNC using log-likelihood. The keywords with the highest log-likelihood values are given in Table 2 below.

4.2 Keyword analysis (Relative frequency analysis)

As mentioned earlier in this paper, linguistic keywords reflect the content of a particular text (Scott, 1997, 2000) through their high frequency.

Table 2 Words with the highest relative and high absolute frequencies

High relative frequency words			High absolute frequency words		
digital	new	data	the	in	as
will	business	services	to	a	
said	year		and	for	
thailand	technology		of	is	

The words with high relative frequencies or keywords are different from the most frequent words, and these keywords reflect the concerns of the *Bangkok Post* news more accurately. These keywords are iteratively categorized into six themes: (1) words relating to digital

technology and its applications, (2) words relating to business and monetary issues, (3) words relating to potential digital technology stakeholder, (4) words relating to digital technology impact, (5) words relating to location and time, and (6) words relating to informational dimension of language (Biber, Conrad & Reppen, 1998). The reliability of these categorizations were rated to have substantial agreement (Cohen's kappa= 0.785). The highest-ranked keywords for each theme with examples of use are given in Table 3 below. This is followed by a discussion of the findings to shed light on the research questions.

Table 3 Keywords categorized by theme (f=Frequency=f; LL= Log-likelihood)

#	Keywords	f	LL	%	Examples
	Words relating to digital technology and its applications				
1	digital	601	+5511.623	32.35%	digital industries, digital infrastructure
2	new	295	+ 168.5181		new board, new age
3	technology	269	+ 1162.164		biometric technology, technology adoption
4	data	248	+ 834.3839		data analysis, data analysis platform
5	services	239	+ 647.7276		online services, service attacks
6	use	194	+ 176.6513		use of cloud services, use cash flow
7	disruption	108	+ 810.706		technology-driven disruption, digital disruption
8	cloud	162	+ 1070.624		cloud and big data analysis, cloud apps and service
9	service	141	+ 210.0264		service channel, service counters attendants
10	mobile	114	+ 729.8153		mobile banking, mobile applications
11	online	108	+ 895.8408		online and mobile, online databases
	Words relating to business activities and monetary issues				
1	business	294	+ 725.9951	29.41%	business activities, business adjusting
2	million	153	+ 306.4581		million baht for research, million baht
3	baht	149	+ 1910.04		baht in revenue, baht devaluation
4	market	145	+ 226.2244		market and scalability, market capitalization
5	billion	133	+ 632.1705		billion baht for state, billion baht annually
6	bank	122	+ 265.9185		bank approved financial institutions, bank by assets
7	financial	119	+ 264.0705		financial burden, financial services
8	banks	114	+ 425.4312		commercial banks, regional banks
9	businesses	107	+ 512.4759		established businesses, businesses like logistics
10	investment	101	+ 268.9911		investment budget, investment products
	Words relating to digital technology potential stakeholder				
1	company	162	+ 208.373	14.71%	distribution company, hi-tech company
2	companies	135	+ 311.6561		global companies, innovative companies
3	government	134	+ 64.52039		government and consumers, government entities
4	industry	132	+ 275.4578		thai tourism industry, advertising industry
5	customers	122	+ 474.9522		new customers, corporate customers
	Words relating to digital technology impact				
1	growth	113	+ 289.1547	8.82%	growth as investment, growth and innovation
2	development	108	+ 109.0896		development agency, development and delivery
3	security	104	+ 242.3782		security awareness, security advisory
	Words relating to location and time				
1	thailand	329	+ 3141.249	8.82%	thailand and indochina, trends in thailand
2	year	283	+ 343.3947		every year, last year
3	country	103	+ 101.2393		leading country, sector in a country
	Words relating to informational dimension of language (Biber, Conrad & Reppen, 1998)				
1	will	553	+264.1475	5.88%	will allow, will aggressively promote
2	said	517	+ 357.053		chief executive said, mr poramin said

The first five themes relate to the digital information chosen by the publisher to be presented to the public through the online version of the newspaper. These themes are dominated by words relating to digital technology and its applications (33.35%) as well as words relating to business and monetary issues (29.41%). So, in the collected online news, it is common to see

words such as *digital industries, digital infrastructure biometric technology, technology adoption, cloud and big data analysis, cloud apps and service, commercial banks, regional banks financial burden, financial services*, *growth as investment, growth and innovation*, *distribution company, hi-tech company* etc. These two themes provide clear indications of the content of the news items presented to the public.

They are followed by words relating to digital technology potential stakeholder (14.71%) and words relating to digital technology impact (8.82%). These words are less frequent than the first two sets despite the massive impact of digital technology to end users. Thus, it can be argued that the publisher prefer to highlight information that are directly related to digital disruption rather than the *after-effect* of its applications. However, given the percentages of the words relating to digital technology and its applications (33.35%) as well as words relating to business and monetary issues (29.41%), the words relating to digital technology potential stakeholder (14.71%) and words relating to digital technology impact (8.82%) are given less importance or quite undermined.

Words relating to location and time (8.82%) and words relating to informational dimension of language (5.88%) show the setting of the news, the people behind the news, and the future perspectives of digital disruption in Thailand, respectively. It is interesting to note the sources cited in the news, but due space constraints, this will be included in a future research study.

5. Conclusion

The ultimate goal of this research study is to investigate what information relating to the phenomenon of digital disruption is presented to the public by The Bangkok Post online newspaper. The main findings showed that information presented on the online version of the newspaper is dominated by (1) words relating to digital technology and its applications, (2) words relating to business activities and monetary issues, and (3) words relating to digital technology potential stakeholder with much less information is presented on digital disruptions potential impacts. While, it is important that the public is provided with the recent information in digital technology such as the phenomenon of digital disruption, it is equally indispensable that the public is warned by providing potential impacts the phenomenon may have on the stakeholders. At the present state of the news items presented on the online version of the newspaper, this is quite far from reality, which may have some repercussions on the reception of information regarding digital disruption.

The main limitation of this research study is its limited data and focus mainly on keywords. Future research studies on digital disruption may examine diachronic data taken from the onset of the Digital Era or a comparative study of digital disruption news may also be conducted. It is also interesting to look at the sources of information cited by the news writers. Most importantly, potential impacts to the reading public may also be investigated.

While others rely on information from the digital disruptors themselves or from the companies that were affected by digital disruption, we hope that this corpus-based investigation of newspaper articles provides another perspective on how an on-going phenomenon of social significance, such as digital disruption, can be examined and understood.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., Hardie, A. and McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press: Edinburgh.

Bradley, C., & O'Toole, C. (2016). *An incumbent's guide to digital disruption*. *McKinsey Quarterly*, *52*(3), 76-85.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bradley, J., Loucks, J., Macaulay, J., Noronha, A., & Wade, M. (2015). Digital vortex: How digital disruption is redefining industries. *Global Center for Digital Business Transformation: An IMD and Cisco initiative*.

Bughin, J. (2017). The best response to digital disruption. *MIT Sloan Management Review*, *58*(4).

Carreon, J. R., & Todd, R. W. (2013). Conflicts between prioritizing medical care and profit-making for a Thai hospital. *International Journal of Language Studies*, *7*(1).

Christensen, C. M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Christensen, C. M. (2013). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Downes, L., & Mui, C. (1998). Unleashing the killer app: digital strategies for market. *Harvard Business School Press, Cambridge, MA*.

Fung, C. C., Khera, V., Depickere, A., Tantatsanawong, P., & Boonbrahm, P. (2008). Raising information security awareness in digital ecosystem with games-a pilot study in Thailand. In *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on* (pp. 375-380). IEEE.

Garrety, K., McLoughlin, I., Wilson, R., Zelle, G., & Martin, M. (2014). *National electronic health records and the digital disruption of moral orders*. *Social Science & Medicine*, 101, 70-77.

Gilbert, R. J. (2015). E-books: A tale of digital disruption. *Journal of Economic Perspectives*, *29*(3), 165-84.

Hongladarom, S. (2016). Cultural politics of the digital divide in Thailand. *วารสาร สำนักบัณฑิต อาสา สมัคร วิทยาลัย พัฒน ศาสตร์ ป๋ ว ย อึ๊ ง ภากร ณ์ มหาวิทยาลัย ธรรมศาสตร์ 13*(1).

Hirt, M., & Willmott, P. (2014). Strategic principles for competing in the digital age. *McKinsey Quarterly*, *5*, 1.

Iordanova, D., & Cunningham, S. (2012). *Digital disruption: Cinema moves on-line*. St Andrews film studies.

Iordanova, D. (2012). Digital disruption: Technological innovation and global film circulation. *HRVATSKI FILMSKI LJETOPIS*, *18*(72), 33-53.

Karimi, J., & Walter, Z. (2015). The role of dynamic capabilities in responding to digital disruption: A factor-based study of the newspaper industry. *Journal of Management Information Systems*, *32*(1), 39-81.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

McCoy, B. R. (2016). Digital Distractions in the Classroom Phase II: Student Classroom Use of Digital Devices for Non-Class Related Purposes. *Journal of Media Education, 7*(1), p. 5-32.

McQuivey, J. (2013). *Digital disruption: Unleashing the next wave of innovation*.

Nickalls, R. (2011). How definite are we about articles in English: A study of L2 learners' English article interlanguage during a university presessional English course. In *Proceedings from the 2011 Corpus Linguistics Conference*.

Quirk, R., Greenbaum, S., Leech, G.N. and Svartvik, J. (1995). A Comprehensive Grammar of the English Language, London: Longman.

Rayson, P., & Garside, R. (2000, October). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora* (pp. 1-6). Association for Computational Linguistics.

Schmidt, E., & Cohen, J. (2010). The Digital Disruption-Connectivity and the Diffusion of Power. *Foreign Aff.*, *89*, 75.

Scott, M. (1997). PC analysis of key words—and key key words. *System*, *25*(2), 233-245.

Statista (2018). Thailand: Which types of publications do you read? Available [Online]: https://www.statista.com/statistics/563404/thailand-types-of-publications-read/ on 14 March 2018.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Thongtep, W. & Pratruangkrai, P. (2016 Oct. 19). "Newspapers covering HM's death become collector's items". *The Nation*. Available [Online]: http://www.nationmultimedia.com/news/business/EconomyAndTourism/30297906 on 14 March 2018.

Weber, M. S., & Monge, P. R. (2017). Industries in turmoil: Driving transformation during periods of disruption. *Communication Research*, *44*(2), 147-176.

Yu, D., & Hang, C. C. (2010). A reflective review of disruptive innovation theory. *International journal of management reviews*, *12*(4), 435-452.

Authors' Biodata

Jonathan Rante Carreon is the Associate Dean of the Faculty of Liberal Arts and the Head of the International Relations of Huachiew Chalermprakiet University and holds PhD in Applied Linguistics and PhD Linguistics. His research interests include, (Critical) Discourse Analysis of data from online sources, health science and formal and informal classroom settings.

Banjob Piyamat is a full-time lecturer in the Department of English, Liberal Arts Faculty at Huachiew Chalermprakiet University and holds M.A. in Linguistics. His research interests include, Applied Linguistics, Discourse Analysis and English Language Teaching.