Twitter for Research, Handbook 2015-2016

Clément Levallois, Morgane Marchand, Tiago Mata, André Panisson

Twitter Mix Days 2015

Edited by Clément Levallois, Morgane Marchand, Tiago Mata, André Panisson

Conference hosted in Lyon, France
by EMLYON Business School,
22-24 April 2015

Foreword

Close to 2,500 research papers on Twitter have been published since 2006, the year the social networking service was founded (see Fausto & Aventurier, chapter 1 in this volume). Scientists have developed an interest in Twitter as an object of study itself, and also as a rich and convenient data source to conduct research on topics independent of Twitter – from the prediction of election results to new challenges in sentiment analysis. The volume and diversity of these scientific contributions would suggest that fruitful exchanges would take place between researchers sharing an interest in Twitter.

Yet, science being structured in separate fields, a consequence is that one researcher using Twitter data in her scientific domain might never cross the path of another scientist using the same type of data but in a different scientific domain. Specialists of business-to-business marketing rarely interact with students in literary studies, while sociologists of urban life seldom engage with the forefront of text mining research –even if they all use Twitter as an input to explore their respective research question. It is reasonable to think that scholars situated in vastly different research traditions would benefit from meeting and exchanging views on how they use Twitter, for which purpose and for what results.

The conference « Twitter for Research » was organized on April 24, 2015 in Lyon, France with the ambition to fulfill this promise: gathering scientists using Twitter in their research, from all disciplinary backgrounds, for a day of exchange, communication and networking. The present volume gathers 13 contributions which were presented this day (see the full program of the conference at `http://tinyurl.com/twitter4research2015`).

This Handbook is the first edition (2015/2016) of an annual series devoted to cover research contributions harnessing Twitter as a key input. To be considered for the next edition, we encourage you to submit a proposal for the next conference "Twitter for Research", due to be organized in Galway, Dublin on April 18-20, 2016.

Clement Levallois
@seinecle
levallois@em-lyon.com

Organization

Executive Committee

Conference Chair	Clément Levallois, Assistant Professor, EMLYON Business School Morgane Marchand, Research Engineer at eXenSa Tiago Mata, Lecturer in Science and Technology Studies, University College London André Panisson, Researcher at the Data Science Laboratory of the Institute for Scientific Interchange in Turin, Italy

Sponsoring Institutions

EMLYON Business School, Lyon-Ecully, France.

Acknowledgements

We would like to thank all the participants for contributing to the success of the conference with a varied program and lively exchanges in panels and in informal settings. We extend our special gratitude to Dr. Jeff Kolb, Data Scientist at Twitter, for delivering a keynote in the spirit of the conference: interdisciplinary in scope, opening broad horizons.

Capgemini Consulting and all our sponsors also receive our sincere thanks for contributing their expertise to the conference and the 2 days preceding it (the "Twitter Mix Days"): DiscoverText, La Cuisine du Web, Learn Assembly, OnlyLyon, Synomia, VisiBrain, FrenchWeb.fr, Social Media Club France.

The events and communication team at EMLYON (Typhaine Grisard, Nadia Brakta, Mathieu Cottereau and Alex Panizzo) made sure that the conference run efficiently, and we are also thankful to them for organizing this day.

Finally, we would like to express our gratitude to Mark Callahan, Senior Platform Growth Manager at Twitter, who helped this initiative take off the ground and develop into a successful event. Thank you.

Clément Levallois
Morgane Marchand
Tiago Mata
André Panisson

Table of Contents

Scientific literature on Twitter as a subject research: Findings based on Bibliometric Analysis

Sibele Fausto[1] and Pascal Aventurier[2]

[1] University of São Paulo (USP)
São Paulo, Brazil
sifausto@usp.br *

[2] Institut National de la Recherche Agronomique, Délégation Information Scientifique et Technique, Services Aéconcentrés for Services Déconcentrés (INRA-DIST-SDAR)
France
Pascal.Aventurier@paca.inra.fr

Abstract. Since its launch in 2006, Internet platform Twitter has rapidly expanded. As a phenomenon of the digital era, Twitter generates a new type of research data that has received a good deal of attention in the academic literature. It has turned into a popular subject research that has been widely investigated in the academic world in different fields ranging from the Social Sciences to Health Sciences, addressing various questions, methods approaches, and covering multiple data sets. This study provides some findings of a bibliometric study which was conducted to describe the scientific literature available on Twitter with descriptive, quantitative information and also in a qualitative approach, in addition to the previous studies and designed as a contribution to a broader picture of how the evolution of the current scientific literature about Twitter is related to bibliographic data sets. Results show a variety of findings that can provide a better comprehension of this social media platform which evolved from a data source for the research to, nowadays, being a research subject itself.

Keywords: Twitter, scientific literature, bibliometric analysis

1 Introduction

Since its launch in 2006, founded by Jack Dorsey and associates in San Francisco, California, USA, Internet platform Twitter (https://twitter.com) has rapidly expanded, "conceived as part of a long line of squawk media, dispatch, short messaging, as well as citizen communications services" (Rogers, 2014), with Dorsey trying to define Twitter as a new medium in itself, a public instant messaging system, meaning also a public information utility (Sarno, 2009).

* Correspondence concerning this article should be addressed to Sibele Fausto, Technical Department of the Integrated Library System, University of São Paulo, Rua da Biblioteca, s/n, Complexo Brasiliana, São Paulo, SP, CEP 05508-050 (Brazil). Email: sifausto@usp.br.

As a phenomenon of the digital era, Twitter generates a new type of research data related to this role as public information utility, emerging as an important source of the new digital paradigm, and has received a good deal of attention in the academic literature. It has turned into a popular research subject that has been widely investigated in the academic world in different fields, ranging from the Social Sciences to the Health Sciences, addressing various questions, method approaches, and covering multiple data sets, with early studies focusing on characterizing tweet types, determining how many of them are of value, and evaluating Twitter as, more or less, of interesting content (Rogers, 2014).

An analysis of the coverage of Twitter's first three years, performed by Arceneaux and Schmitz Weiss (2010), pointed out the prevalence of "food tweets" and the more general "mindless stream", emanating from Twitter in its early years, leading Rogers (2014) to ask what value lies in breakfast and lunch tweets, and responding that "geo-located food tweets may be of interest to those studying the geography of taste and other questions of cultural preference" (p. xiii-xiv).

Thus, together with other social web platforms, such as Facebook, blogs, Wikipedia, YouTube, and others, Twitter becomes a new source of information for many researchers in different disciplines, providing a variety of data to extract (Metaxas & Mustafaraj, 2014). Despite being behind services like Facebook in numbers (with 1.3 billion active users) and WhatsApp (500 million), this platform currently has about 284 million active users worldwide, with 500 million tweets being sent out every day in more than 30 different languages (Twitter, 2015). In 2014 the eMarketer Consultancy predicted that Twitter should reach 300 million monthly active users by 2016, and pass the 400 million mark by the end of 2018.

Since its launch Twitter has incorporated growing facilities for sharing content, with options to add links, images and videos, and efficient mechanisms to aggregate specific topics by hashtags: labels assigned to a term, preceded by the hash symbol (#), making the term an active link that allows recovery of all tweets with the hashtag. Hashtags not only gather and record the tweets on the same subject but also make it possible to recover its track record, through monitoring or mining. Moreover, Twitter's dynamics, where users develop a virtual community made up of a public identified by common interests - followers, who share and replicate content freely with each other in an intense interaction through actions, such as the "re-tweet" (RT), generating affiliations around themes, topics and identities (Zappavigna, 2011). All these characteristics make Twitter a major data source to track the performance of personal and institutional microblogs, communities and the most diverse subjects through systems that monitor the interactions of real-time messages generated on the platform, targeting to various analyses, hence the growth of scientific literature on Twitter.

Bruns and Weller (2014) pointed out that scientific publications, with "Twitter" explicitly in the title, are as follows: 1,400 on Scopus® database, 470 on Web of Science® (WoS) database and another 10,000 available via Google Scholar. In turn, Zimmer and Proferes (2014) conducted a content analysis study with a set of 382 academic publications that used Twitter from 2006 to 2012. Also, Weller et al (2014) published a comprehensive study on Twitter, addressing aspects of this social media related to its concepts and methods, even as

its perspectives and practices in use in society, as those from popular culture, brand and crisis communication, politics and activism, journalism, and in academia.

This study provides some more insights in this matter through a bibliometric study which was conducted to describe the scientific literature available on Twitter, with quantitative information and also in a qualitative approach, in addition to the previous studies and designed as a contribution to a broader picture of how the evolution of the current scientific literature about Twitter is related to bibliographic data sets.

2 Methods

The methodological approach chosen for this study is Bibliometric Analysis, an original area of study generally covering books and publications and whose principle, according to Tague-Sutcliffe (1992), is the analysis of scientific or technical activity through quantitative studies of publications. Bibliometric Analysis provides insights on which stage a given research area lies, and it is a methodology also used in several areas, with the objective of characterizing the published research output and the forecasting of trends, yielding results that are also useful to support decision-making (Rostaing, 1996).

Through Bibliometric Analysis it is possible to study some quantitative parameters of the publications, periodicals, authors, keywords, users and citations (Pao, 1989). But nowadays Bibliometric Analysis integrates the broader area of metric studies of the information, with different scopes and finalities in the discovery of information characteristics in general, not just about science, comprising and combining different methodologies and tools to provide a more comprehensive view of its objects of study (Cronin & Sugimoto, 2014).

This Bibliometric Analysis was carried out for the indexed literature in the period from 2006 (when Twitter was released) until 2014, and only included articles with the word "Twitter" in the title, abstract and/or keywords. Analysis was conducted on the results of the Scopus® database (from Reed-Elsevier), which has a wider coverage than WoS® (from Thomson Reuters). The 2,338 documents retrieved were analyzed on two levels. The first was a basic one related to quantitative indicators, according to the average growth in the number of publications per year, identifying main research areas that have published about Twitter, as well as journal titles, most prolific authors, institutions and countries. This was followed by a second level with a deeper analysis, considering citations and the 'hot topics' of the literature, through the combination of techniques and tools of lexical analysis and content about the corpora, with the software Sphinx Lexica®, that allows the researcher to differentiate in a semi-automatic form the main terms, identifying those that are 'hot topics' in the publications about Twitter, and CorText Manager[3] , that was used to discover the relationships between the different topics and countries involved, providing maps of heterogeneous networks as thematic and collaborative maps, sorted by term frequency per countries and by a chi-squared (χ^2) distribution of the "hot topics".

[3] CorText Manager: urlhttp://www.cortext.net/.

3 Results

3.1 Quantitative indicators

The academic literature with Twitter as research subject has seen an average growth rate of 52.23% per year, from one sole article published in 2006 to the 720 published in 2014. Figure 1 shows that this growth was ongoing and has accelerated after 2009, reaching the peak in 2014. This result points out that in fact Twitter is becoming a consolidated object of research.

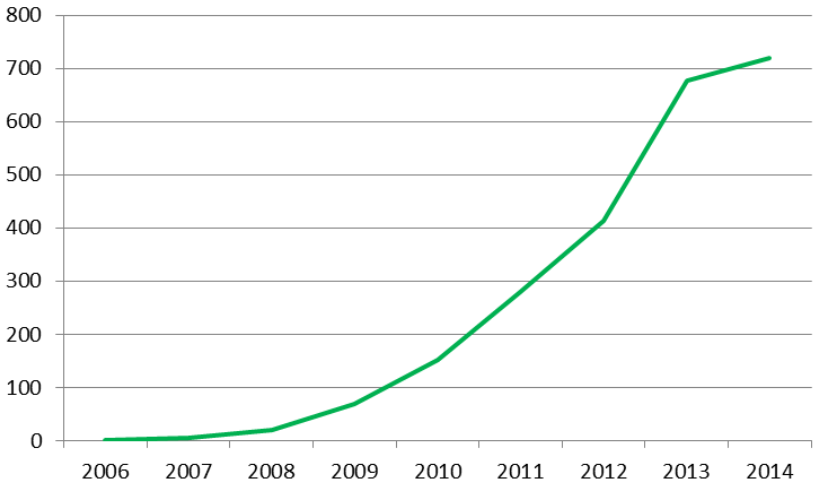


Fig. 1: Growth of the scientific literature about Twitter (2006-2014).

As for the research areas which have published about Twitter, Figure 2 shows the dominance of certain disciplines with more than 100 publications, such as the Social Sciences, with 1,014 articles (33% of the total sample); and Computer Science, with 826 articles (27%), followed by Medicine, with 327 articles (11%). But we can observe that other disciplines also appear with relevant participation, such as Business Management and Accounting, and Engineering, which had 284 and 273 papers, respectively, both reaching 9% of the total sample; followed by Arts and Humanities, with 215 (7%) and Mathematics, with 105 (4%).

Related to journal titles with more than 10 publications in this matter, results were as follows: PLoS One ($n = 46$); Computers in Human Behavior ($n = 37$); Journal of Medical Internet Research ($n = 37$); First Monday ($n = 36$); Information Communication and Society ($n = 35$); Public Relations Review ($n = 34$); *Estudios Sobre El Mensaje Periodistico* ($n = 19$); Social Science Computer Review ($n = 18$); E-content ($n = 17$); Cutting Edge Technologies in Higher Education ($n = 16$); New Media and Society ($n = 15$); Expert Systems with Applications ($n = 14$); ACM Transactions on Intelligent Systems and Technology, Journal of Communication, Government Information Quarterly, and PC World San Francisco CA, 13 articles each; Journal of the American Society for Information Science and Technology, and

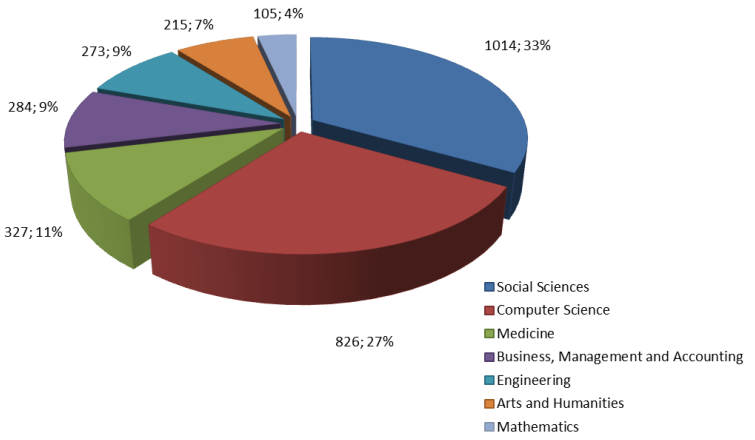


Fig. 2: Number of articles by disciplines.

Online Wilton Connecticut with 12 articles each; American Behavioral Scientist ($n = 11$); Cyberpsychology Behavior and Social Networking, Proceedings of the ASIST Annual Meeting, and Studies in Computational Intelligence, with 10 articles each [Figure 3].

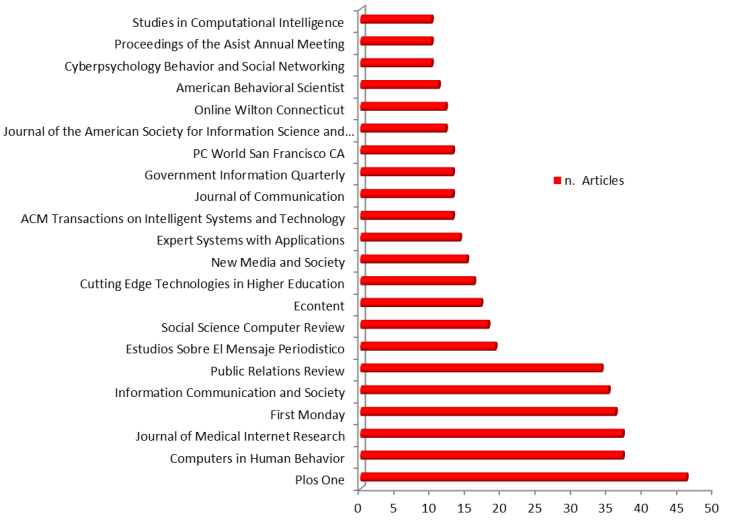


Fig. 3: Top journal titles with articles published about Twitter (2006-2014).

The most prolific authors were H. W. Park with 13 articles and A. Bruns, with 12 ones, but others authors also figured with a good research output in this matter, such as J. J. Jung and T. Highfield, both with 9 articles; P. R. Spence and S. H. Burton, with 8 articles; and A. O. Larsson, I. Himelboim, J. K. Harris and M. Thelwall with 7 articles each [Table 1].

Table 1: Most prolific authors who published about Twitter (2006-2014)

Author	n. Articles
Park, H.W.	13
Bruns, A.	12
Jung, J.J.	9
Highfield, T.	9
Spence, P.R.	8
Burton, S.H.	8
Larsson, A.O.	7
Himelboim, I.	7
Harris, J.K.	7
Thelwall, M.	7

As for institutions with more than 15 contributions, results show Yeungnam University with the most number of articles published (30 contributions); followed by Queensland University of Technology ($n = 22$); Pennsylvania State University, the University of Oxford and the University of Toronto ($n = 18$ each); University of Texas at Austin ($n = 17$); University of Maryland and Indiana University ($n = 16$ each); Seoul National University, University College, (London), Carnegie Mellon University, University of New York State at Buffalo, with 15 articles each [Figure 4].

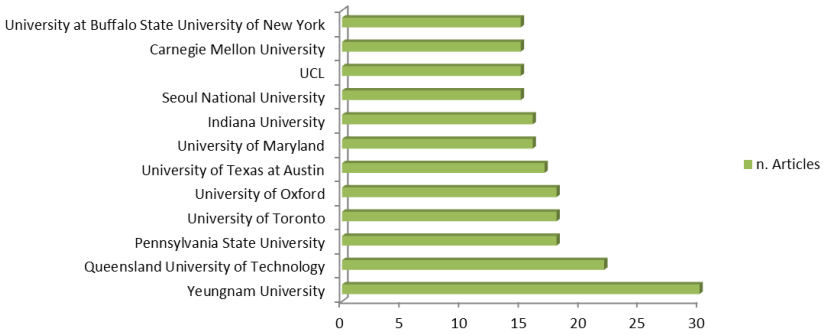


Fig. 4: Top institutions with more than 15 articles published about Twitter (2006-2014).

The following countries had more than 50 articles: the United States (USA) with 912 publications (39% of the total sample); the United Kingdom (UK) with 216 (9.23%); Spain, 148 (6.33%); Australia, 119 (5.1%); South Korea, 116 (4.9%); China, 99 (4.23%); Canada, 98 (4.2%); Germany, 65 (2.8%); Japan, 63 (2.7%); and The Netherlands, 51 (2.2%) [Figure 5].

Ahead, in this analysis, we will show those same countries according to the topics most published by them, forming clusters by equivalent subject.

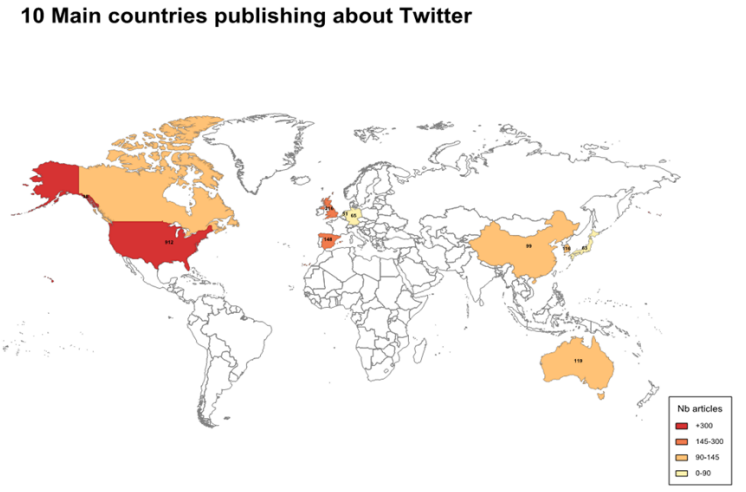


Fig. 5: Top countries with more than 50 articles published about Twitter (2006-2014).

3.2 Citations and textual analysis

Most cited authors and journals. In addition to the basic quantitative indicators shown before, we also examined the data sample to verify the most cited papers in order to get insights about the impact of that scientific literature with Twitter as subject research. Table 2 shows articles with more than 100 citations, summarized by authors, article titles, publication year, journal titles, number of citations received and the topics.

The most cited authors were Kaplan and Haenlein (2010), with an article published in the "Business Horizons" journal, which reached 1,296 citations. As the journal title points out, the main topic of this paper is Business. Next, the authors cited about 400 times were Jansen et al (2009), with an article cited 438 times, in the topic of Information in the "Journal of the American Society for Information Science and Technology;" and Bollen, Mao and Zeng (2011), cited 430 times, with a paper on the topic of Computation, in the " Journal of Computational Science".

The third group of most-cited authors, which reached about 200 citations, were Kietzmann et al (2011) with an article on the topic of Business, again in the "Business Horizons" journal, with 283 citations; Marwick and Boyd (2011) on the topic of Information in "New Media and Society," with 243 citations; and, finally, Yardi, Romero and Schoenebeck (2009), with an article in "First Monday" on the General topic that reached 239 citations.

The fourth and last group of most-cited authors spanned 116 to 176 citations and includes 14 journals, with topics like Science (3 occurrences), Health (5 occurrences), Information (2 occurrences), Computation (2 occurrences) and Business (1 occurrence) [Table 2].

Only four journals with most-cited articles are also in the top journals list with more than 10 articles published about Twitter, as follows: "PLoS One" has published 46 articles on Twitter and also figured among the most-cited with two articles, from Chew and Eysenbach

Table 2: Most prolific authors who published about Twitter (2006-2014)

Author	Title	Year	Journal	n. Citations	Topic
Kaplan A.M.; Haenlein M.	Users of the world, unite! The challenges and opportunities of Social Media	2010	Business Horizons	1296	BUSINESS
Jansen B.J.; Zhang M.; Sobel K.; Chowdury A.	Twitter power: Tweets as electronic word of mouth	2009	Journal of the American Society for Information Science and Technology	438	INFORMATION
Bollen J.; Mao H.; Zeng X.	Twitter mood predicts the stock market	2011	Journal of Computational Science	430	COMPUTATION
Kietzmann J.H.; Hermkens K.; McCarthy I.P.; Silvestre B.S.	Social media? Get serious! Understanding the functional building blocks of social media	2011	Business Horizons	283	BUSINESS
Marwick A.E.; Boyd D.	I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience	2011	New Media and Society	243	INFORMATION
Yardi, S.; Romero, D.; Schoenebeck, G.	Detecting spam in a Twitter network	2009	First Monday	239	GENERAL
Chew C.; Eysenbach G.	Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak	2010	PLoS One	176	SCIENCE
Hawn C.	Report from the field: Take two aspirin and tweet me in the morning: How twitter, facebook, and other social media are reshaping health care."	2009	Health Affairs	170	HEALTH
Boyd D.; Crawford K.	Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	2012	Information Communication and Society	167	INFORMATION
Golder S.A.; Macy M.W.	Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures	2011	Science	161	SCIENCE
Signorini A.; Segre A.M.; Polgreen P.M.	The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic	2011	PLoS One	161	SCIENCE
Scanfeld D.; Scanfeld V.; Larson E.L.	Dissemination of health information through social networks: Twitter and antibiotics	2010	American Journal of Infection Control	147	HEALTH
Junco R.; Heiberger G.; Loken E.	The effect of Twitter on college student engagement and grades	2011	Journal of Computer Assisted Learning	143	COMPUTATION
O'Keeffe G.S.; Clarke-Pearson K.; Mulligan D.A.; et al.	Clinical report - The impact of social media on children, adolescents, and families	2011	Pediatrics	131	HEALTH
Thelwall M.; Buckley K.; Paltoglou G.	Sentiment in Twitter events	2011	Journal of the American Society for Information Science and Technology	131	INFORMATION
Ebner M.; Lienhardt C.; Rohs M.; Meyer I.	Microblogs in Higher Education - A chance to facilitate informal and process-oriented learning?	2010	Computers and Education	131	COMPUTATION
Vance K.; Howe W.; Dellavalle R.P.	Social Internet Sites as a Source of Public Health Information	2009	Dermatologic Clinics	128	HEALTH
Hennig-Thurau T.; Malthouse E.C.; et al.	The impact of new media on customer relationships	2010	Journal of Service Research	121	BUSINESS
Lee Hughes A.; Palen L.	Twitter adoption and use in mass convergence and emergency events	2009	International Journal of Emergency Management	116	HEALTH

(2010) which received 176 citations, and from Signorini et al (2011), with 161 citations. The "First Monday" has published 36 articles on Twitter and received 239 citations in the article by Yardi, Romero and Schoenebeck (2009). The "Information Communication and Society" has published 35 articles and received 167 citations from Boyd and Crawford (2012). Finally, the "New Media and Society" journal published 15 articles on Twitter and received 243 citations to the Marwick and Boyd's article (2011).

So there isn't a clear relationship between the largest number of published articles on Twitter and a greater number of citations received, as only four $(18, 18\%)$ from a total of 22 journals with more than 10 publications on Twitter received more than 100 citations.

"Hot Topics" in the literature about Twitter. Continuing our analysis, we proceeded in the mining of the most frequent topics into literature about Twitter, through applying the software Sphinx Lexica® in a semi-automatic lexical corpus analysis, after removing the term "Twitter" from it, identifying specific terms with more than 100 occurrences by year to observe their evolution in frequency during the analyzed period of 2006-2014. All terms were joined together in groups of equivalent ones.

Results show several most frequent topics, with "social media" leading in a total of 3,197 occurrences, with an ongoing growth in its frequency over the period, followed by the topic "social networks," with the total of 1,981 occurrences, also with an ongoing growth in its frequency over the same period.

"Hot topics" also very often include the topic of "learning" (human learning), with 360 occurrences; "health" (330); "political activity" (324), among others represented in Table 3.

Representative keywords were evolving rapidly. For example, "Second Life" was representative in the early years. In turn "big data" and "data mining" appear over the whole time period. It was observed that in general most of the topics had a continued growth in frequency since the year in which they first appeared in the literature on Twitter, and this evolution of their frequency can predict the trends of these topics.

This topic analysis also allowed the mapping of these topics by clusters, according to related lexical similarities, e.g. a topic cluster related to "health" groups terms such as "public health," "patient" and "health care," (represented in clear blue in Figure 6), as well as a topic cluster related to terms such as "political activity," "social movements," "public relations" and "mass communication" (represented in clear yellow in Figure 6, that also shows the other topic clusters that emerged in this analysis).

This "hot topic" analysis also allowed clustering topics related to their frequency by countries, as shown in Figure 6, and also pointing out the relations between countries by thematic collaborations, next in Figure 7.

In Figure 6 there are countries that are more present in each cluster, addressing related topics sorted by decreased term frequency, such as USA, China, UK, Spain, Australia, Canada, South Korea, Japan, etc. But in Figure 7 it is possible to observe the thematic collaborations between these countries and others that were checked through the institutional affiliations (by country) in the recovered articles, with a chi-squared (χ^2) distribution of the "hot topics". It was observed that there are only 340 papers with international collaborations,

Table 3: "Hot topics" and their evolution in literature about Twitter (2006-2014)

Topics	2006	2007	2008	2009	2010	2011	2012	2013	2014	TOTAL
social_media	0	0	1	23	10	370	494	984	1211	3197
social_networks	0	4	19	50	128	221	373	592	593	1981
learning	0	0	0	1	10	38	61	93	157	360
health	0	0	0	4	11	30	35	122	127	330
political_activity	0	0	0	0	3	38	44	114	125	324
mobile	0	2	2	14	28	32	67	68	69	282
real_time	0	0	6	1	11	28	38	60	53	197
content_analysis	0	0	0	0	10	20	30	58	73	191
sentiment_analysis	0	0	0	0	0	12	25	56	89	182
marketing	0	0	0	5	8	22	34	61	45	175
twitter_users	0	0	1	2	2	11	29	48	69	162
case_study	0	0	0	5	5	11	30	47	57	155
opinion	0	0	0	0	0	13	23	45	66	147
public_health	0	0	0	4	5	18	21	40	50	138
facebook	0	0	0	3	9	24	17	25	48	126
news	0	0	1	1	4	22	24	32	41	125
qualitative_analysis_data	0	0	0	0	3	18	19	24	54	118
web_sites	0	0	1	4	19	19	23	25	23	114
big_data	0	0	0	0	0	0	27	30	55	112
large_scale	0	0	0	0	6	7	23	26	43	105
higher_education	0	0	0	1	10	18	16	36	22	103
user_activities	0	0	0	0	1	2	19	28	49	102
second_life	0	1	0	1	5	10	0	2	0	19

13% of the whole corpus, which is a very low percentage compared to other domains in international research.

These clusters, for example, show such themes as "collective action", "big data", "text and web mining", "marketing", "political activity" and "user generated content" and were common topics in the collaborations between Brazil, France, India, Italy, UK, and others. In turn, topics such as "crisis communication", "corporate communication", "traditional education", "political activity" and "sentiment classification", were most common in collaborations between Belgium, Switzerland, The Netherlands, and Turkey. This analysis is useful to realize which countries have collaborations in the research of specific topics about Twitter.

4 Discussion and concluding remarks

These findings indicate that there was a steady growth in the academic literature which includes Twitter as research subject in the period from its founding, in 2006, to 2014. The subject areas with most contributions were the Social Sciences, followed by Computer Science and Medicine. There is a long tail of journal titles that have published articles on this subject, with 22 titles in more than 10 publications, all of them in English, apart from a single title in

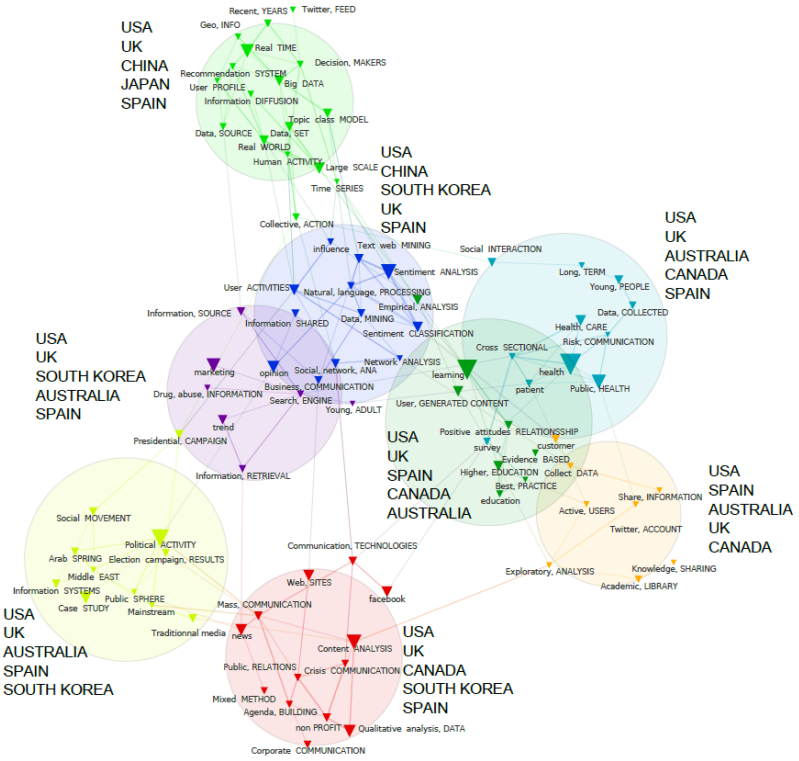


Fig. 6: “Hot topics” by countries sorted by decreased term frequency.

Spanish. Two authors stand out in terms of the number of articles, and various institutions have made contributions, most of them being universities from North America (USA and Canada) and Europe (UK, Spain, Germany and The Netherlands), although Asia (China, Japan and South Korea) and Oceania (Australia) also stood out.

With a second-level analysis in a deeper manner, it was possible also consider citations and the “hot topics” of the literature about Twitter, in order to verify their impact, and to identify which topics were more discussed in this literature, as well as to find out with a mapping tool how these topics were distributed by subject groups and countries, including collaborations between these countries. It was observed that there is a group of papers which received more than 100 citations, but there was not a clear relation between the largest number of published articles on Twitter and a greater number of citations received.

There are several “hot topics” that were addressed in the literature on Twitter, with “social media” and “social networks” leading with 3,197 and 1,981 occurrences, respectively, and the majority of these topics had a continued growth in their frequency in the literature during the analyzed period. Twitter analysis is strongly related to social issues and enhanced problems linked to youth populations, as health (obesity), public health, international disasters, drug abuse, but also to the learning and education process. Three other important fields of analysis are social behavior (opinion), political activity and branding/media strategy.

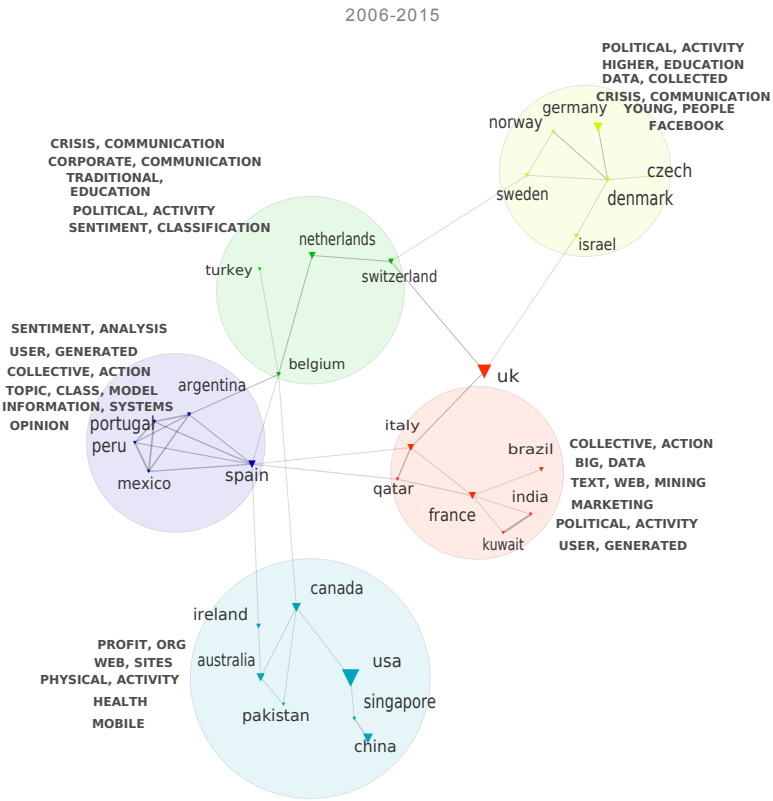


Fig. 7: "Hot topics" by country collaborations sorted by χ^2 distribution.

It was interesting to see how these "hot topics" were organized under common subjects and how they were distributed between different countries, including collaborations between these countries in the research of certain topics about Twitter. There are a few international collaborations (13% of the corpus) and it is mostly delimited to traditional collaborator countries (between USA, UK, and those from Asia – China and South Korea).

All these findings add to and group together with previous studies about Twitter and aim at a contribution to a better comprehension of this social media platform which evolved from a data source for the research to, nowadays, being a research subject itself.

Future directions of the investigation of this role of Twitter as an object of study through Bibliometric Analysis are a continuous task that point to other approaches which can provide more detailed insights, such as how the research network is drawn from the cited references.

Bibliography

[1] Arceneaux, N., & Schmitz Weiss, A. (2010). Seems stupid until you try it: Press coverage of Twitter, 2006–9. *New Media & Society*, 12(8), 1262–1279. doi:10.1177/1461444809360773.

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2 (1), 1-8. doi:10.1016/j.jocs.2010.12.007.

[3] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679. doi:10.1080/1369118X.2012.678878.

[4] Bruns, A., & Weller, K. (2014). Twitter Data Analytics – Or: the Pleasures and Perils of Studying Twitter. *Aslib Journal of Information Management*, 66 (3). DOI: http://dx.doi.org/10.1108/AJIM-02-2014-0027.

[5] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, 5(11), e14118. doi:10.1371/journal.pone.0014118.

[6] Cronin, B., & Sugimoto, C. R. (Eds.) (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press. 466 p.

[7] eMarketer Consultancy (2014). *Emerging Markets Drive Twitter User Growth Worldwide*. Accessed 9 Jun. 2015 from: `http://www.emarketer.com/Article/Emerging-Markets-Drive-Twitter-User-Growth-Worldwide/1010874\#sthash.pxPo8DI3.dpuf`.

[8] Jansen, B.J., Zhang, M., Sobel, K., & Chowdury A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60 (11), 2169-2188. doi:10.1002/asi.21149.

[9] Kaplan A.M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. doi:10.1016/j.bushor.2009.09.003.

[10] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241-251. doi: 0.1016/j.bushor.2011.01.005.

[11] Marwick, A. E. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114-133. doi:10.1177/1461444810365313.

[12] Metaxas, P. & Mustafaraj, E. (2014). Sifting the sand on the river bank: Social media as a source for research data. *IT-Information Technology*, 56 (5), 230–239. doi:10.1515/itit-2014-1047.

[13] Pao, M. L. (1989). *Concepts of information retrieval*. Englewood, Colorado: Libraries Unlimited, Inc. 285 p.

[14] Rogers, (2014). Debanalising Twitter: The Transformation of an Object of Study. In Weller, K., Bruns, A., Burguess, J., Mahrt, M., & Puschmann, C. (Eds). *Twitter and Society* (pp. ix- xxvi). New York: Peter Lang Publishing.

[15] Rostaing, H. (1996). *La bibliométrie et ses techniques.* Marseille: CRRM.135 p.

[16] Sarno, (2009). Twitter creator Jack Dorsey illuminates the site's founding document. Part I. *Los Angeles Times*. Accessed 15 Jun. 2015 from: `http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html`.

[17] Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One*, 6(5), e19467. doi:10.1371/journal.pone.0019467.

[18] Tague-Sutcliffe (1992). An Introduction to Informetrics. *Information Processing Management*, 28 (1), 1-3. doi:10.1016/0306-4573(92)90087-G.

[19] Twitter Inc. (2015). *About*. Accessed 9 Jun. 2015 from: `https://about.twitter.com/company`.

[20] Weller, K., Bruns, A., Burguess, J., Mahrt, M., & Puschmann, C. (Eds)(2014). *Twitter and Society*. New York: Peter Lang Publishing. 486 p.

[21] Yardi, S., Romero, D., & Schoenebeck, G. (2009). Detecting spam in a twitter network. *First Monday*, 15(1). doi: http://dx.doi.org/10.5210/fm.v15i1.2793.

[22] Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*, 3(5), 788-806. doi:10.1177/1461444810385097.

[23] Zimmer, M. & Proferes, N. J. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66 (3), 250-261. doi: 10.1108/AJIM-09-2013-0083.

Elements for an epistemology of instrumentation and collaboration in Twitter data research

Eglantine Schmitt

Sorbonne Universités, Université de Technologie de Compiègne
Proxem, Compiègne
France
eglantine.schmitt@utc.fr

Abstract. Twitter has become the most studied online social network in academia, in social sciences as well as in other fields. It is commonly grasped through a collection and analysis of its own data. In this paper, I show through a bibliometric analysis that scholarly publications on this matter come equally from social and computer sciences, and from natural sciences to a lesser extent. Social scientists rely mostly on classical quantitative methods while computer scientists try to improve algorithms and techniques. Twitter data can take several epistemic values, from representing nothing to representing real-world social phenomena. Having observed the infrequence of interdisciplinary works, I make a few suggestions based on the history of science for future collaborative projects based on Twitter data.

Keywords: Twitter, epistemology, bibliometrics, instrument research, interdisciplinarity

1 Introduction

In a few years' time, Twitter has become the most studied online social network (OSN) in academia. For Tufekci (2014), it is to Web study what the *Drosophilia melanogaster* is to biology: a model organism, chosen for convenience at first, then studied so much more than any other social network that it polarises research, giving scholars a well-documented reference where one can add cumulative knowledge. The choice of a model organism is never entirely arbitrary; there are usually good reasons to choose one in particular, but those may be more practical than scientific ones. In Twitter's case, the fact that most tweets are public, added to the availability of an application programming interface (API) which remained more or less stable over time, makes sufficient amounts of Twitter data available to researchers, which is basically the condition of possibility of any scientific investigation on the matter. This is probably why Twitter is, as of today, much more studied than Facebook, although the latter has much more users (1.441 billion versus 302 million monthly active users during Q1 2015, according to Statista 2015) and should then provide a richer and more detailed view of online sociability. What sets Twitter and Facebook apart, from a data collection point of view, is that, while Facebook also provides an API, it only shows the private data that one is allowed to see, that is, posts and comments from one's friends. As a consequence, full data sets analysis is possible only to Facebook R& D team and their potential partners, any

result based on Facebook data would be otherwise a result on a researcher's own Facebook friends' data.

While Twitter data can be easy to collect for a knowledgeable user, Twitter in itself is not just any online social network, it is one of a kind. It is a singular sociotechnical object that will not fit in a standard definition, from a media and communication point of view. Kwak, Lee et al. (2011) in particular, have shown that it is both a tool for sociability and information, hence a medium and a social network, and maybe fundamentally neither of the two. From a linguistics perspective, it can be seen as a repository of oral or journalistic language (Rogers 2013), a language of its own kind, or a mix of community 'dialects' (Bryden, Funk and Jansen 2013). Just like the Drosophilia melanogaster, what can be said of Twitter may not be true of any other online social network, nor networking in general. It would then be crucial to researchers to be well aware of the theoretical requirements for any generalisation from Twitter data analysis.

Describing and mapping a specific digital object is usually the role of social sciences and the humanities, and especially media and communication studies. As a new OSN arises, social scientists may want to explore whether the notion of sociability evolves through it, and this is made possible by Twitter since followers/followings and mentions form empirical networks which can actually be retrieved and visualized. In the same way, linguists would study how people talk or write on Twitter and compare those idiosyncrasies to general languages. From an epistemological point of view, this is regular, usual, social science and humanities, or, to phrase it in Kuhnian terms, normal science.

However, new kinds of data analysis based on Twitter have emerged from many other scientific fields related to computer science, machine learning and so on, with a much more systematic media coverage. According to those approaches, it is made possible to predict the success of Kickstarter campaigns (Etter, Grossglauser and Thiran 2013), box-offices (Asur and Huberman 2010) or the stock market (Bollen et al. 2011). With the required conceptualisation from the social sciences and humanities, we might be witnessing the emergence of a new scientific field and new ways to analyse society and language, that we might call 'computational sociology' (Hummon and Fararo 1995), 'computational social science' (Lazer, Pentland et al. 2009) or 'social physics' (Pentland 2014).

While articles claiming they can predict a social phenomenon through Twitter data analysis have been heavily press-covered, what matters to the study of the evolution of science is how much of that kind of work is actually published and how it is considered by other scientists. Computational study of Twitter might be completely anecdotal to less press-covered 'normal' social science, but may also be about to replace it. Objectifying and quantifying the balance between computer sciences and social sciences in the study of Twitter is the purpose of this paper.

In this respect, I studied how Twitter is used in academia, what it is aimed for and which theoretical status is conferred to Twitter data. I did so by collecting all scientific papers about Twitter for 6 months (a total of almost 300 papers), and registering for each one the field in which authors work, the subject of the paper, the methods they use and some other things such as the paper source (journal or conference), the researcher's university's country, the

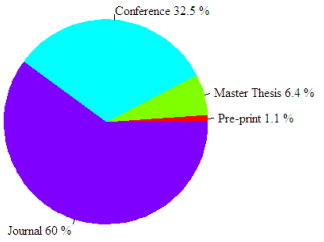


Fig. 1: Publication by format.

language in which they write and the publication date (Part 1). Through this bibliometric study, I thus empirically demonstrate that, in the quantitative picture, research based on Twitter (or research about Twitter) equally comes from computer sciences on the one hand, and social sciences and the humanities on the other hand, with a lesser contribution from natural sciences (Part 2). With that in mind, I discuss what 'studying Twitter' actually means and what signification can be given to Twitter data. In the light of this reflection, I suggest how the quality of research based on Twitter data analysis might be improved through actual interdisciplinary collaborations (Part 3).

2 Bibliometric data collection

A list of papers was constituted through the keyword request 'twitter' using Google Scholar alerting system for six months. Other publication sources such as arXiv or PLOS One are known to be rather focused on certain fields, such as computer science or biology, and would have introduced a bias in my data collection since it was not meant to be focused on a specific field. Also, Google Scholar tends to reference fringe journals and conferences which might be under-represented in well-established scientific repositories. As more innovative approaches tend to appear on the frontiers of science rather than in core journals, the choice of Google Scholar was also a way to include in my set more of the 'computational social science' mentioned above. The keyword itself was non-ambiguous and hence usually brought good results, but publications not mentioning Twitter in the title or abstract were not included in the data set because of the method; however, while an empirical proof of that has not been made, I could not find a reason why this sample would not be representative of research about Twitter as a whole.

Data collection took place from July 26th, 2013 to January 27th, 2014. 282 articles were found, including 60% of journal article and 32.5% of conference papers. The remaining 7.5% constituted in Master's theses and pre-prints. [Figure 1] For each article, I manually retrieved the title, authors' names, publication date, university's country, journal or conference name and field, language used, and subject and method(s) used according to the abstract. Coding the scientific field and method was the most delicate task as there is a larger part left to subjectivity and no standard definition of what a field is. For a more accurate coding, I used a three-level typology consisting of 1) field as described by journal's or conference's name,

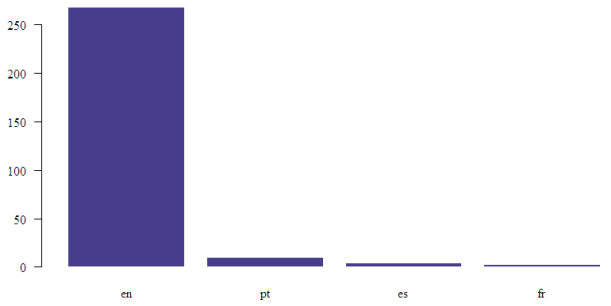


Fig. 2: Publication by language.

author's lab or a similar information source, which was 2) grouped in field families and finally 3) in 'meta fields' which were computer sciences, social sciences and natural sciences. To code the method, I collected techniques and elements of methodology that were explicitly mentioned in the title or abstract with very little rephrasing, hence promoting papers which considered that the method mattered enough to appear in the abstract. While coding the method, I hid the scientific field column of my data set so that I would not be influenced or tempted to align certain methods with certain fields. It is worth mentioning that the distinction between tweet collection and tweet sampling can be difficult to make; as a rule of thumb I considered that, when the abstract suggested that tweets were manually annotated, the authors had probably worked on a sample of maybe hundreds of thousands of tweets rather than a set of millions of tweets. When several methods or algorithms were mentioned, I took them all; when nothing was mentioned, I left the field empty rather than projecting assumptions on a paper. On the whole, less than ten papers did not show any method at all.

3 Data analysis

A search on papers related to Twitter cannot be considered representative of academia in general; however, it provides geographical and linguistic information about this fragment of scientific research, which may or may not resemble the whole scholarly world. In our case, English is, by far and as expected, the most used language in the scientific community; other represented languages are Portuguese, Spanish and French. [Figure 2] In terms of country of affiliation, the United States is obviously the most productive; then come European and Asian countries. [Figure 3] The latter seem to be particularly active as far as computer sciences are concerned. This is rather consistent with the geography of science in general.

In terms of scientific fields, media and communication studies are the most represented, followed by social sciences in the broad sense and knowledge engineering. Rather new fields such as digital studies and social network analysis (SNA) also have quite a prominent place. Some less expected matters such as health or education seem to have a solid interest in studying Twitter. [Figure 4] When gathered in meta-fields as described above, subjects are quite well balanced between social sciences (46.1%) and computer sciences (41.8%), what remains being grouped in natural sciences (12.1%). [Figure 5] The latter are mainly

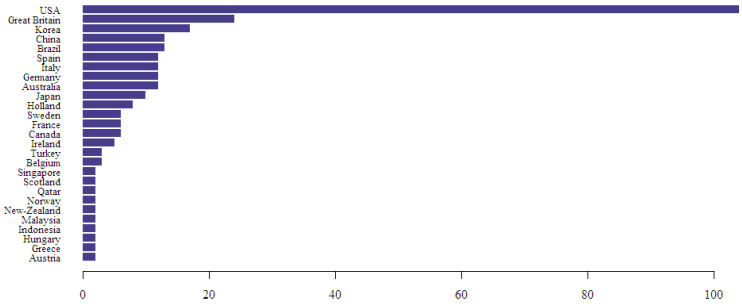


Fig. 3: Publication by researcher affiliation country.

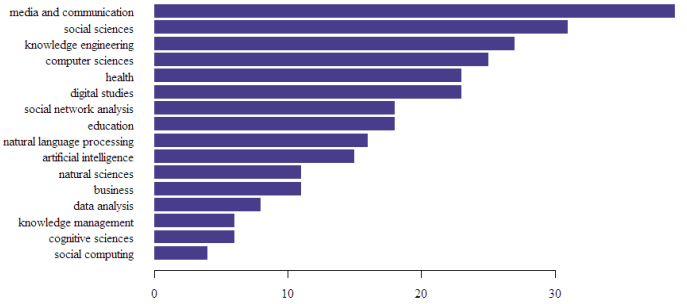


Fig. 4: Publication by field.

represented by health scientists and professionals wondering how Twitter can be used for a better understanding of the general public, epidemics detection and scientific communication, and physicists trying to apply the mathematical methods from their field to the analysis of Twitter data. In each of those meta fields, the conference/journal ratio can be quite different. Twitter research in social and natural science is much more published in journals, while computer scientists seem to write in journals and conferences in a most balanced way. [Figure 6] Researchers from those three fields informally confirmed to me that journals are more prestigious in natural and social sciences while computer scientists prefer to publish in conference proceedings. A way to interpret this distribution is to consider that, Twitter research being aligned with publication habits, Twitter as a subject is (already) a rather prestigious one.

However, my purpose what not to determine whether Twitter is studied in A-list journals but how scientific practices have taken a grasp on the subject. To this end, one of the main results of this paper is to provide a weighted list of the methods used and mentioned by researchers in their attempts to explore this rather new subject. Overall, collecting or sampling are the most frequently noted techniques, which is quite logical since any data analysis such as social network analysis or text mining relies on data acquisition. [Figure 7] More surprisingly, writing an essay or surveying people without collecting tweets seems to be have remained quite a frequent way to address such a contemporary question; those

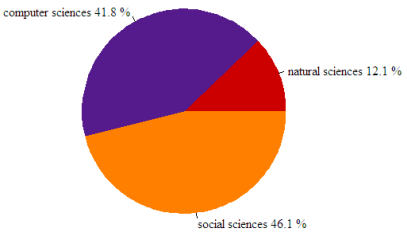


Fig. 5: Publication by meta field.

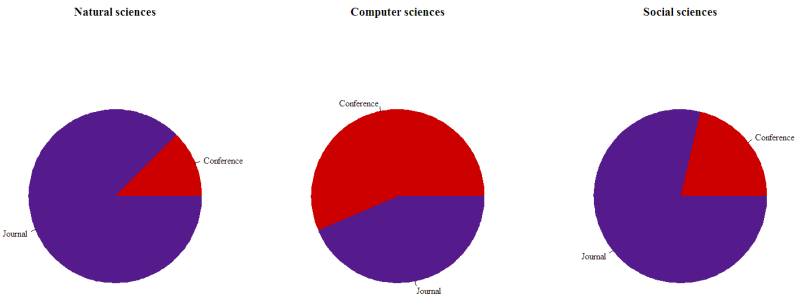


Fig. 6

papers are about Twitter but they are not based on actual Twitter data. This is especially true in social sciences since other fields scarcely resort to those classical methods. [Figure 8] Social scientists' approach can be quantitative, relying on a statistical analysis of some variables related to Twitter, but they are hardly computational, with the exception of social network analysis, which has become a prominent tool in the social sciences since Moreno's seminal (and non-computational) work in 1934. Social network analysis (SNA) is also quite noticeable in computer sciences, amongst other text and data mining techniques such as time series analysis or Latent Dirichlet Allocation. [Figure 9] There seems to be no essays about Twitter written by computer scientists, who rely on data collection and analysis when contributing to academic research. When confronting those techniques to the researchers' goal as presented in titles and abstracts, we observe that Twitter is basically used to provide data for regular mining tasks such as classification, event detection or sentiment analysis. Natural sciences resemble both other fields, with a strong presence of collection and sampling, but also a long tail of computational techniques [Figure 10].

In a nutshell, we can say as of now that there are two major attitudes towards Twitter in research:

- in social sciences, classical methods are called up to address a new research subject;
- in computer sciences, more modern techniques are applied on Twitter data to attain classical goals.

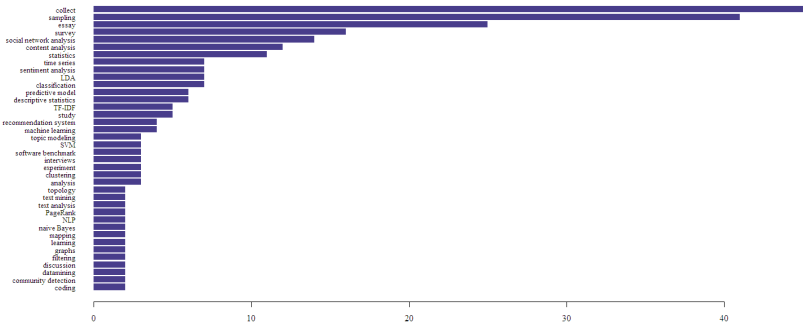


Fig. 7: Publication by method (appearing more than once).

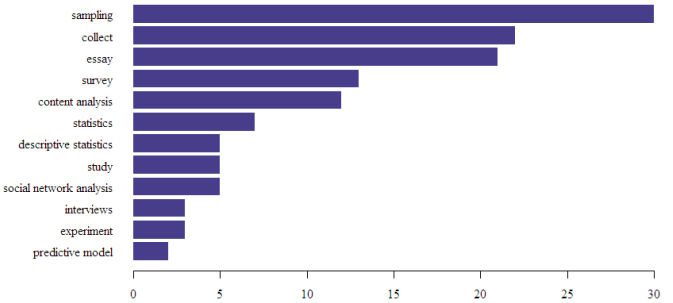


Fig. 8: Methods in Social Sciences (appearing more than once).

A closer observation of the publications' meta-data suggests that there are, at best, very few exchanged views between those attitudes, or in other words, very little interdisciplinary work between computer scientists and social scientists. It seems then that the observations of C.P. Snow (1990) regarding the two cultures of natural sciences and the humanities can be played again to describe the (non-) connections between computer and social scientists in the context of Twitter research.

In addition to a methodological comparison of scientific fields, we can adopt a more epistemological view of Twitter research. Through a more qualitative approach, I tracked how authors define the meaning they confer to Twitter data, or in other words, what they believe Twitter data to be a representation of. A selection of those definitions appears on Table 1. What stands out in those definitions is that Twitter data can have several epistemic values, depending on the field, subject and stakes of the paper, namely:

(1) Twitter data can be analysed for a better knowledge of Twitter itself or the way people behave, interact or communicate on Twitter;
(2) Twitter data can serve as any other data from any other source, to fuel the calibration of an algorithm, without consideration of what those data may represent;
(3) Twitter data can be used for a more ambitious purpose, where it is a proxy for real-world phenomena (such as stock markets or elections) as shown in Table 1.

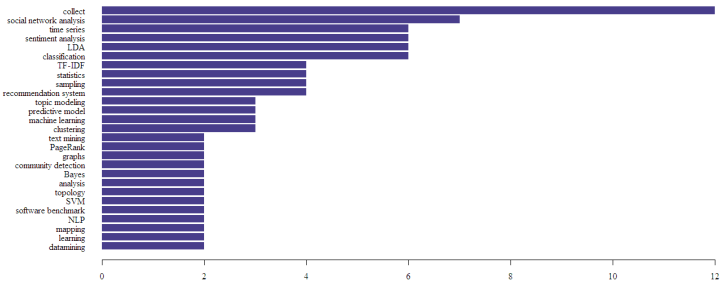


Fig. 9: Methods in Computer Sciences (appearing more than once).

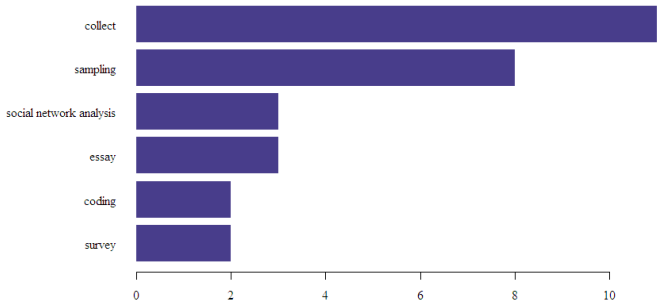


Fig. 10: Methods in Natural Sciences (appearing more than once).

Several issues emerge from this typology. First, none of those points are unquestionable and each of them should probably be discussed independently within each affected field. Second, not all papers explicitly define which epistemic value they confer to Twitter and it sometimes changes within the paper itself. Lastly, some papers which are not written by social scientists make assumptions on social phenomena from Twitter or the 'real world': I recognize here what one may call the 'hubris of big data', where someone with computer skills begins to make assumptions belonging to social sciences, claiming for instance that they are able to predict the future through Twitter data (the truth of that claim being a different matter). For that matter, the epistemic values of Twitter as described above are somehow correlated to a field, but there are many exceptions to each case. However, as can logically be inferred from what has just been suggested, we can say with confidence that the 'hubris of big data' is more typical of computer sciences than social sciences.

4 Improving research driven by Twitter data

Social and computer sciences have shown different agendas in Twitter research, the former trying to understand what Twitter is and how it modifies what sociability is, the latter aiming at better algorithms that would accurately classify information and eventually predict phenomena. What is also happening is that new kinds of knowledge about social phenomena are emerging outside social sciences, anchored in what a social scientist would call the instrumentation of her scientific practice. In other words, computer sciences no longer play

Table 1

Quote	**Reference**
'The recent advent of social networking sites and the data within such sites now allow researchers to empirically validate these theories'	Cha and Haddadi 2010
'data from the social media services Foursquare and Twitter often reflects activities of users at real-world venues including restaurants, coffee shops, shopping venues and more.'	Grinberg et al. 2013
'The use of information gathered via Twitter messaging may provide a method to overcome some of these difficulties, allowing for the assessment of mood states in a set of users through analysing the contents of their textual communications in real time and on a large scale.'	Lampos et al. 2013
'Internet-based activities have demonstrated to be highly reflective of real-world situations.'	Ostrowski
'Although the Twitter stream contains much useless chatter, by virtue of the sheer number of tweets, it will still contain enough useful information for tracking or even forecasting behavior when extracted in an appropriate manner.'	Signorini et al. 2011
'Since the majority of tweets are conversational in nature, they form a good source for public opinion. As a result of its large, diverse, and growing user base, Twitter has emerged as an important source for online opinion and sentiment indexes.'	Hassan et al. 2013

an ancillary role for sciences and seem to have turned into something that is neither entirely a science yet nor really a technique any more. Through the case of Twitter research, we are possibly witnessing the emergence of a new scientific paradigm, or at least, a reconfiguration of the interrelation between the tools and the object of a given science.

While this reconfiguration is new indeed in the case of Twitter analysis, and of what we may loosely call 'data sciences', this is not the first one as regards the history of sciences and especially social sciences. In natural language processing, a field that was born and gradually freed itself from linguistics, it is a common saying that computer systems dedicated to language analysis improve every time a linguist is eliminated. In the last decade, Aurélien Bénel (2013), a researcher with a computer science background, has been looking for ways to improve interdisciplinary collaborations between engineers and social scientists where computer science is not purely an ancillary way to build a tool for research, but a field of interdisciplinary research and experiment in itself. Grossetti and Boë (2008) studied the history of speech analysis and how engineers grasped the subject by taking the question of speech in the broadest sense outside the field of phonetics and language sciences in general. At first, they were working for linguists but soon began to work with some open-minded linguists such as phonologist René Gsell, and finally became an independent research area. As a matter of fact, Terry Shinn (2008) has argued that this kind of research, neither purely

scientific nor purely academic, called 'instrument research', has existed in fact ever since the birth of modern sciences in the 17th century.

In my view, what happened to linguistics and phonetics in the 20th century is currently happening to social sciences in a broad sense. As Aurélien Bénel pointed out, social scientists have better options than simply using computer science as a tool just like digital humanists mostly did until now. The design and construction of computational tools for the collection and analysis of digital data can be done through an actual collaboration between computer and social scientists, without a sponsor-provider relationship and starting from the moment research goals are set. More precisely, the success of the collaboration should be one of those goals. I believe that, through this configuration, research based on Twitter data and digital data as a whole can be improved with an actual conceptualisation of both the object of the research (the social phenomena) and the tools to reach it and build new knowledge.

5 Conclusion

In this paper I have shown that Twitter has become an equally popular source of data for both social and computer scientists. While their works are fed from the same source, they pursue very different objectives, some of them being typical of their field, some more innovative. On the one hand, social scientists conceptualise Twitter itself and try to understand how it transforms our definition of sociability; on the other hand, computer scientists design tools to analyse those data and make assumptions about real-world social phenomena. Twitter data can therefore have several epistemic values, which are not systematically expressed by the researchers themselves; all in all, everything points to the emergence of new grounds and methods to study social phenomena. From my point of view, this evolution looks like what happened to language sciences in the 20th century, which means that researchers from all fields have the possibility to take inspiration from the history of science and improve their scientific practices and results through interdisciplinary instrument research.

Bibliography

[1] Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. In 2010 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 492–499. IEEE.

[2] Bénel, A. (2013). Quelle interdisciplinarité pour les « humanités numériques » ? *Les Cahiers Du Numérique*, 1, 1–23. doi:10.3166/LCN.9.1.25-38

[3] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

[4] Bryden, J., Funk, S., & Jansen, V. (2013). Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, 1–9.

[5] Cha, M., & Haddadi, H. (2010). Measuring user influence in twitter: The million follower fallacy. *Proceedings of International AAAI Conference on Weblogs and Social.*

[6] Etter, V., Grossglauser, M., & Thiran, P. (2013). Launch hard or go home! In *Proceedings of the first ACM conference on Online social networks - COSN '13,* pp. 177–182. New York, New York, USA: ACM Press.

[7] Grinberg, N., Naaman, M., Shaw, B., & Lotan, G. (2013). Extracting Diurnal Patterns of Real World Activity from Social Media.

[8] Grossetti, M., & Boë, L.-J. (2008). Sciences humaines et recherche instrumentale: qui instrumente qui? *Revue d'Anthropologie des Connaissances*, 2, 1(1), 97.

[9] Hassan, A., Abbasi, A., & Zeng, D. (2013). Twitter Sentiment Analysis: A Bootstrap Ensemble Framework. *Proceedings of the ASE/IEEE.*

[10] Hummon, N. P., & Fararo, T. J. (1995). The emergence of computational sociology. *The Journal of Mathematical Sociology*, 20(2-3), 79–87.

[11] Lampos, V., Lansdall-Welfare, T., Araya, R., & Cristianini, N. (2013). Analysing Mood Patterns in the United Kingdom through Twitter Content. *arXiv Preprint*, 1–8.

[12] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Van Alstyne, M. (2009). Social science. Computational social science. *Science* (New York, N.Y.), 323(5915), 721–3.

[13] Ostrowski, D. A. (n.d.). Identification of Trends in Consumer Behavior through Social Media.

[14] Moreno, J. (1934). *Who shall survive? A new approach to the problem of human interrelations.* Washington, DC: Nervous and Mental Diseas Publishing Co.

[15] Pentland A. (2014). *Social Physics: How Good Ideas Spread-The Lessons from a New Science.* Penguin Press.

[16] Shinn, T. & Ragouet, P. (2000). Formes de division du travail scientifique et convergence intellectuelle. La recherche technico-instrumentale. *Revue Française de Sociologie*, 41(3), 447–473.

[17] Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS One*, 6(5), e19467.

[18] Snow, C. P. (1990). The Two Cultures. *Leonardo*, 23(2), 169–173.

[19] Statista, "Number of monthly active Facebook users worldwide as of 1st quarter 2015 (in millions)", May 2015, online: `http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/` (checked on May 19th 2015)

[20] Statista, "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2015 (in millions)", May 2015, online: `http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/` (checked on May 19th 2015)

[21] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.*

Text mining and Twitter to analyze British swearing habits

Michael Gauthier[1], Adrien Guille[2], Fabien Rico[3], and Anthony Deseille[4]

[1] CRTT, University of Lyon 2
Lyon, France
michael.gauthier@univ-lyon2.fr
[2] ERIC, University of Lyon 2
Lyon, France
adrien.guille@univ-lyon2.fr
[3] ERIC, University of Lyon 1
Lyon, France
[4] University of Lyon 1
Lyon, France

Abstract. The way women and men speak and are expected to behave is frequently discussed. For example, women are sometimes described as speaking more than men, and men as swearing more than women. These stereotypes can alter people's expectations concerning the way we should behave. Indeed, if the idea that females generally swear less frequently than males is widespread, women who swear may be perceived as deviant from the norm, and thus be stigmatized. Clearly understanding what is true and what is not in these studies and reports is not an easy task, because there is a considerable amount of differing opinions on the topic. The way swear words are used by women and men is one of those topics which remains vague, but whose stake is great, since swearing is often considered as an act of power and a way of affirming oneself.

This article will introduce the data gathered from a corpus of tweets in order to shed a new light on new ways of analyzing specific sociolinguistic features like gendered uses of swear words on Twitter. Analyzing the linguistic behavior of users of these media can be an interesting way of generating a most contemporary corpus representative of general trends, and computational linguistics can represent a very accurate and powerful method of analyzing the different uses people can make of certain speech patterns. In order to carry out the study, we used several tools taken from both computer science and linguistics. These tools may represent innovative methods to analyze the effect of social parameters on speech patterns displayed in Twitter corpora. Thanks to this data, we analyze both quantitative, and qualitative instances of swear words in the corpus, to see how the linguistic gendered preferences may differ when swearing is used, but just as importantly, we see how comparable they can be. Indeed, very often when dealing with gender in corpus linguistics, small differences tend to be focused on, whereas they are actually minor compared to the similarities. As for every study, the methods used here also have certain limits that we present as well. Without pretending to be representative of interactions other than the computer-mediated ones present in this corpus, we hope that this data can shed an up-to-date

and neutral light on the way women and men use swear words on Twitter, and on the implications these results may have, as well as on new tools researchers can use in various areas of research. We believe that this study can also be useful to computational linguists/sociologists thanks to the methods used to access data not directly available and displayed by users (e.g. the age or the sex).

Keywords: Linguistics, swear words, gender

1 Introduction

The way women and men speak is a common source of discussions and debates, be it in academic research or in mainstream media. Many social attitudes and linguistic features have generally been attributed either to women or men; swearing is one of those topics traditionally associated with men. As Coates (2004) reported, "*the folklinguistic belief that men swear more than women and use more taboo words is widespread*", consequently leading to the creation of stigmas preventing women or men from using a certain type of linguistic behavior without being stigmatized. These preconceived ideas also fuel societal stereotypes and may impact people's standards concerning what is desirable from each gender. Moreover, swearing is often considered as an act of power (see Beers Fägersten, 2012; Murray, 2012; Lakoff, 2004; G. I. Hughes, 2006) and a way of affirming oneself. Thus, the fact that one gender may be perceived as more frequent users of swear words, or on the other hand as swear words eschewers, may have an impact on other qualities related to power we would inherently attribute to one gender or the other, whether or not these differences are real. A certain number of studies have shown that contrary to what has long been widely believed, women do not swear less than men, nor do they use a drastically different register (Baruch & Jenkins, 2007; Hammons, 2012; S. E. Hughes, 1992; Jay, 1992; Baker, 2014; Thelwall, 2008; Coates, 2004). Indeed, some of these surveys have shown that what generally differs between women's and men's use of swear words is not the rate at which they are used, but the context in which they are used, as well as the kinds of words women and men use. Some studies seem to indicate that the use of "strong" swear words (see below for a description of what "strong" swearing is) by women is increasing in certain contexts, and especially on social media (Murray, 2012; Thelwall, 2008), and in the United Kingdom. Thelwall predicted that "*a reversal in gender patterns for strong swearing will slowly become more widespread, at least in social network sites*", and this seemed especially true for younger generations of users (users aged 16-19 in the case of Thelwall). Without pretending that our article would answer this question, this is a good example to illustrate how social media can be one way of highlighting new linguistic features, and try to better understand them. The importance and time we devote to social media sites is growing every year according to a study from Ofcom (see the 2013 Ofcom report), and it concerns people from all age groups, and all socioeconomic backgrounds (Smith & Brewer, 2012). It would also seem that on these media, and especially on Twitter, people tend to swear more than in face to face interactions (Wenbo et al., 2014). This paper will introduce the data gathered from a sample corpus of about one million tweets from about 18,000 users collected through Twitter in order to see

how data such as age and gender can be put in relation with more (socio)linguistic-related features on Twitter, and how these features can be analyzed. Here, the goal of this pilot study is more experimental and demonstrative, as we will try to suggest innovative ways of empirically analyzing gender and language, rather than give definitive answers. In order to carry out our investigation, we collected tweets localized in the UK, then we analyzed the corpus with various tools (event detection programs, lemmatizers, named-entity recognition etc). Information concerning the users (sex, age and location) was retrieved so that we could draw statistics and study different social, contextual and linguistic uses of swear words.

Thanks to this data, we will analyze both quantitative, and qualitative instances of swear words in the corpus, to see how linguistic gendered preferences may differ when we swear, but just as importantly, we will see how comparable they can be. As Baker (2014) pointed out, in a lot of studies dealing with gender in corpus linguistics, small differences often tend to be focused on, whereas they remain minor compared to the similarities, thus giving the erroneous idea that these represent proofs of an inherent divergence between sexes. To try to avoid that, this paper will try to remain as neutral as possible, to see whether a clear distinction can be made out of this corpus or not. As for every study, the methods used here have certain limits that we will discuss here as well.

2 Corpus description

When carrying out a linguistic study of any kind, the researcher needs a corpus on which they will base their analyses. The way this corpus is collected will directly impact the kind of data gathered, as well as the implications, limitations, and the reliability of this data. Thus, the corpus collection phase has to be carefully thought to have results matching the different aspects the researcher wants to investigate.

2.1 What is swearing?

As mentioned in the introduction, certain studies seem to indicate that strong swearing among women may be becoming more and more common on social media. However, being able to determine what can be considered a swear word, and whether it is offensive or not is not an easy task, as not everyone has the same latitude regarding swearing. Indeed, children who swear will sometimes be severely reprimanded, whereas it may go unnoticed among adults (Ladegaard, 2004). Also, people's own perceptions of swear words may influence how offended they are by them (Jay, 1992; Stapleton, 2010), and thus not everyone will be offended by the same words. The way swearing is perceived varies a lot between generations for example (Harris, 1990), and some people may not even consider that certain words are swear words, whereas others will (S. E. Hughes, 1992). For these reasons, it is hard to define a clear and empirical list of swear words on which everyone will agree. In order to compile a list of swear words that would be as objective and appropriate to our sample as possible, we needed to have a standard, a list of words considered swear words by most British speakers. In this regard, the study from Wenbo et al. (2014) seemed to be a good start, as they have managed to put together a list of 788 English swear words and their variations. These words

were manually and independently annotated by two native speakers of English who both agreed that these words are "mostly used for cursing". We decided to use the Wenbo et al. (2014) study as a standard on which we would base certain aspects of the methodology and analyses of our pilot study, because we deemed that their study can be considered as a reference point for our own investigation partly due to the fact that their research was carried out in 2014, so it is to this day one of the most contemporary on this topic. It is also very exhaustive, as their corpus is composed of 51M English tweets from all around the world, so their results are more likely to be representative of global trends on Twitter. One of the conclusions they came to is that of the 788 words they used to define swearing tweets, *"the top seven swear words - fuck, shit, ass, bitch, nigga, hell and whore cover 90.40% of all the curse word occurrences"* in their corpus. These seven words alone then represent the vast majority of the swear words repertoire of Twitter users in their sample.

Considering the afore-mentioned fact, and in order to be able to maximize the relevance of what we can consider as a swearing tweet in our sample, we chose to include the 20 most common swear words in the Wenbo et al. study. This alone should ensure a reliable representativeness of what can be used to differentiate a swearing tweet from a non-swearing one. However, it could be argued that this list may not necessarily be representative of a majority of native speakers of English, especially as it was sampled by only two native speakers. We mentioned earlier the fact that people's attitudes regarding swear words tend to vary a lot, so relying on two people only seems limited to be able to build a comprehensive list of swear words, especially when considering that for the sake of our study, we wished to focus on the UK only. Thus, what was significant in Wenbo et al.'s study may not be as significant in our own sample, as their study was based on a sample of the worldwide stream of tweets from a given period, whereas our corpus is much more localized, so there may also exist a geographical bias. In order to limit those bias as much as possible we also used the swear words mentioned in the editorial guidelines concerning the use of offensive language by the British Broadcasting Corporation (BBC) and which were not present in the list taken from Wenbo et al.. The BBC can be considered as representative of a standard in terms of what should be labelled as a swear word in the UK, especially as this concerns what is acceptable or not from audiences[5]. We deemed that this would represent a reliable addition we could use to create a list of widely accepted offensive words applicable to a British sample. In the end, the complete list of words we used to distinguish swearing tweets from the others is composed of *fuck, shit, ass, bitch, nigga, hell, whore, dick, piss, pussy, slut, tit, fag, damn, cunt, cum, cock, retard, blowjob, wanker, bastard, prick, bollocks, bloody, crap, bugger*. In other words, if a tweet contains any one, or more, of these words, it will be considered as a swearing tweet.

[5] For more details on the guidelines regarding what the BBC considers as offensive language, see: `http://www.bbc.co.uk/guidelines/editorialguidelines/advice/offensivelanguage/index.shtml`

2.2 Social media and swearing

It is inadvisable to presume that the speech patterns displayed on social media are accurately representative of trends present in face to face conversations, especially on Twitter, as users are faced with a limit of 140 characters which does not apply in face to face interactions. However, it can be interesting to compare the way people swear on social media to the way they swear in oral contexts to be able to better understand how these two modes of communication can be compared, and how representative of "real life" trends swearing on Twitter can be. As we discussed earlier, the time dedicated to social media sites like Twitter increases every year. In 2013, a study from Ofcom[6] revealed that in 2012 in the UK, a vast majority of people from all age groups and socio-economic backgrounds used social media. A majority of these people also reported using social media more than once a day, which was not the case in 2011. This illustrates the growing importance that the Internet, and social media in particular, are gaining in our daily lives, consequently increasing the likelihood of daily speech patterns and evolutions of certain linguistic attitudes being present on social media and vice versa. According to Wenbo et al., "one out of 13 tweets contains curse words" (Wenbo et al., 2014). As corpora of tweets can be composed of a virtually unlimited number of tweets, the proportion of potentially interesting swear words to analyze thereby represents a very appealing way of generating data.

As we stated earlier, the way swear words or other linguistic resources are used and perceived by a specific community can have an impact on the way this community is considered. Conversely, the way swear words are used inside a group can be an indication of evolutions in the way this community identifies itself with regards to its status, or its power for example (Beers Fägersten, 2012; Lakoff, 2004; Murray, 2012; G. I. Hughes, 2006). Since women from the United Kingdom seemed to be the most likely to use strong swear words more than men in previous studies (Thelwall, 2008), we figured that studying the use of swear words of British women and men on Twitter may reveal more profound changes in people's perception and use of swear words, at least in online communities present on Twitter.

3 Methodology

The main requirements we had in order to be able to carry out our study were thus the age of the informants (as younger generations of women seemed to be the most likely to experience this increase in swearing on social media sites), their gender, and they had to be localized in the UK, as this region seemed to be the most sensitive to the aforementioned phenomenon. Twitter's API (Application Programming Interface) seemed to be the perfect solution in this regard, as it can offer access to every one of these parameters. In order to collect our corpus, we used the streaming API and only requested tweets from the United Kingdom by mentioning the corresponding geolocation. We let the collection of tweets run between 7 April and 15 May 2015 and got a total number of 961,186 tweets from 18,060 users.

[6] See the Ofcom report on Adults' media use and attitudes report, 2013.

3.1 Inferring gender on Twitter

Users' gender is determined thanks to the name they provided. We created two repositories of female and male names given to British babies since the 1950s[7], and they are composed of a total of about 30,000 gendered names. For every user whose tweet we collect, the program automatically checks whether the name provided is present in one file or the other (i.e., whether the name is present in the male or the female repository), and if it is present in one, and not the other, the user is attributed the corresponding gender. As a way to avoid any bias with ambiguous names (names which can be given both to women or men, like Robin for example), if the name is present in both files, the user is considered as undefined, and is rejected.

3.2 Inferring age

The age is determined thanks to the information provided by users in the description of their profiles. We have defined a list of patterns which allows the program to automatically identify a digit sequence in a description that corresponds to the age of the user (e.g. thanks to regular expressions, the program will identify 25 from "I'm 25 yo" for example). In order to maximize the accuracy of our results, we decided to split users according to their gender and age groups. We thus categorized users into six different age groups which will be referred to as described in Table 1.

Table 1: Table of notations

Notation	Meaning
$\mathcal{C}^m_{12-18}$	Tweets published by males aged 12-18
$\mathcal{C}^m_{19-30}$	Tweets published by males aged 19-30
$\mathcal{C}^m_{31-45}$	Tweets published by males aged 31-45
$\mathcal{C}^m_{46-60}$	Tweets published by males aged 46-60
$\mathcal{C}^f_{12-18}$	Tweets published by females aged 12-18
$\mathcal{C}^f_{19-30}$	Tweets published by females aged 19-30
$\mathcal{C}^f_{31-45}$	Tweets published by females aged 31-45
$\mathcal{C}^f_{46-60}$	Tweets published by females aged 46-60

The reason why we chose those age groups is because as many sociolinguistic studies have shown, people we spend a lot of time with can have an influence on the way we speak, especially among children (Eckert, 2008; Stapleton, 2010; Ladegaard, 2004). Since children spend most of their time at school, with peers of the same age, children of the same educational level are more likely to display similar speech patterns. Thus, until age 18, users

[7] Sources: General Register Office, National Records of Scotland and Office for National Statistics.

are classified according to the academic level they are the most likely to belong to in the United Kingdom.

According to the Office for National Statistics, in 2013 the average age of mothers was 30 in England and Wales[8], so age 30 will be used as a marker for two age groups. Indeed, studies suggest that parents who have children produce more standard forms than usual and avoid the use of taboo language (Stapleton, 2003; Mercury, 1995), so having babies is likely to influence the linguistic attitudes of people from these generations, hence the need to take it into account in our age classification. These age groups should allow the heterogeneousness of our sub-corpora to be limited as much as possible. Such groups also have the advantage of limiting the interference of problems caused by users who may not keep their profiles up to date for example, and who may claim to be 22 in their descriptions, whereas they would now be 23. The age reported would in this case not be the actual age of the user, but they would still belong to the most appropriate age group.

4 Analysis and results

To help us analyze this vast collection of tweets and gain insights into the contexts in which Twitter users swear, we leverage several data analysis tools developed in the field of machine learning. But first, some classic data concerning the demographics of our corpus will help us understand its composition.

4.1 Distribution of the number of tweets per gender and age

Table 2 and Figure 1 present basic data about the demographics of our corpus. Unsurprisingly, as shown in Figure 2, there is a huge imbalance in the representation of the different age groups taken into account, with a vast majority of our users reported as being between 12 and 30 years old. This was to be expected and this repartition also corresponds to the most represented age groups on Twitter as a whole.

Also, what a manual verification revealed is that for both the youngest and the oldest age groups (i.e. the 5-11 and the 61-99), most of the users' descriptions do not correspond to actual human users, or are representative of anomalous profiles, like pages dedicated to companies or pets (for the youngest age group), or clearly untrustworthy profiles (for the oldest age group). This is mainly due to the method we used to gather information concerning the age of our users, based on regular expressions, and which does not make a difference between a profile dedicated to a 4 year old dog, and a 4 year old boy, although a 5 year old boy is unlikely to have a Twitter profile. Our regular expressions are meant to look for profiles mentioning a number followed by "years old" in users' descriptions (or variations of "years old", like "yo", as it is a very common way to mention one's age on Twitter), among others, but does not take into account any mention of gender, or of being a human, as this is only processed thanks to the name provided. So, to prevent the potential interference of this dubious data, these age groups are never taken into account in the analyses we make, and are just presented here out of a concern for transparency.

[8] See the 2014 report from the Office for National Statistics.

Table 2: Basic corpus properties

	Male	Female	Total
# of users	10313	7747	18060
# of tweets	579864	381322	961186

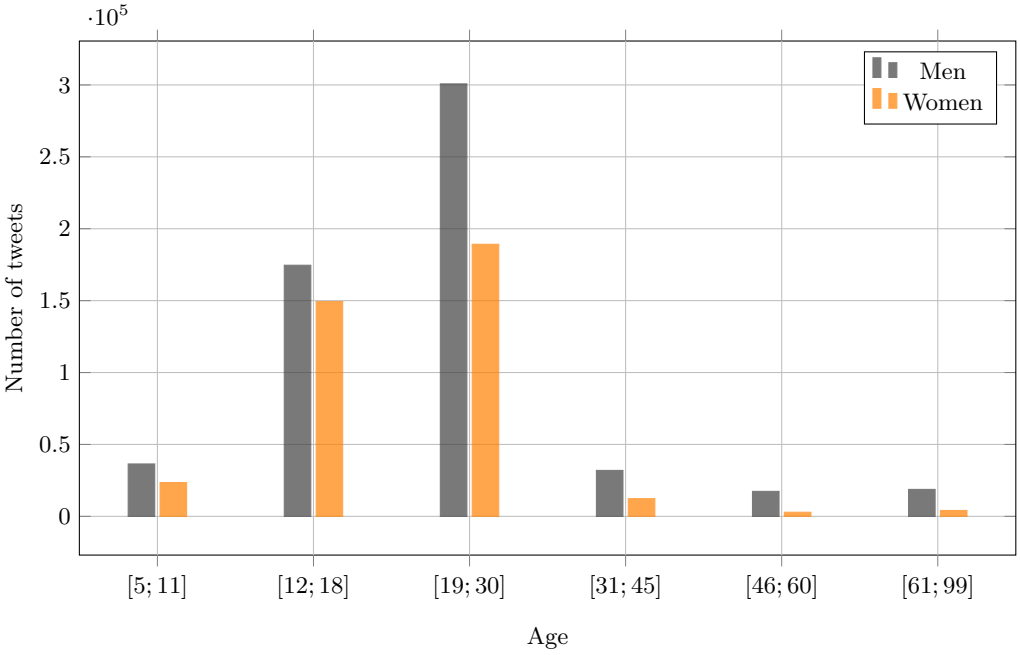


Fig. 1: Distribution of the number of tweets per gender and age.

4.2 Distribution of the number of tweets per user

Figure 2 plots the distribution of the number of tweets per user on a log-log scale. We note that it follows a power-law, with one user contributing over 3700 tweets in the corpus, while a lot of users contribute to fewer than 100 tweets, thus showing that the interference of potential spam accounts producing a great number of tweets in a short amount of time is very limited.

4.3 Proportion of swearing tweets among women and men

In our corpus, 5.8% of the male tweets contained at least one swear word, compared to 4.8% for women. Figures 3 and 4 present the proportion of tweets containing the eleven most common swear words for women and men. However, as percentages of this kind do not provide much information about the specific use of each word, we normalized the frequency of each swear word on one million words for both women and men. The results are presented below in Table 3.

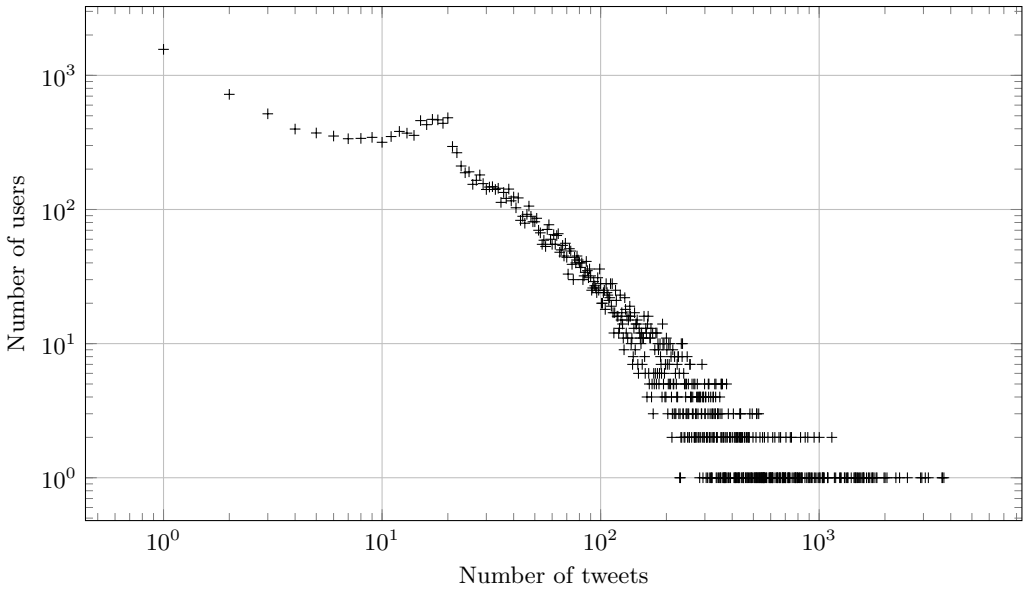


Fig. 2: Distribution of the number of tweets per user on a log-log scale.

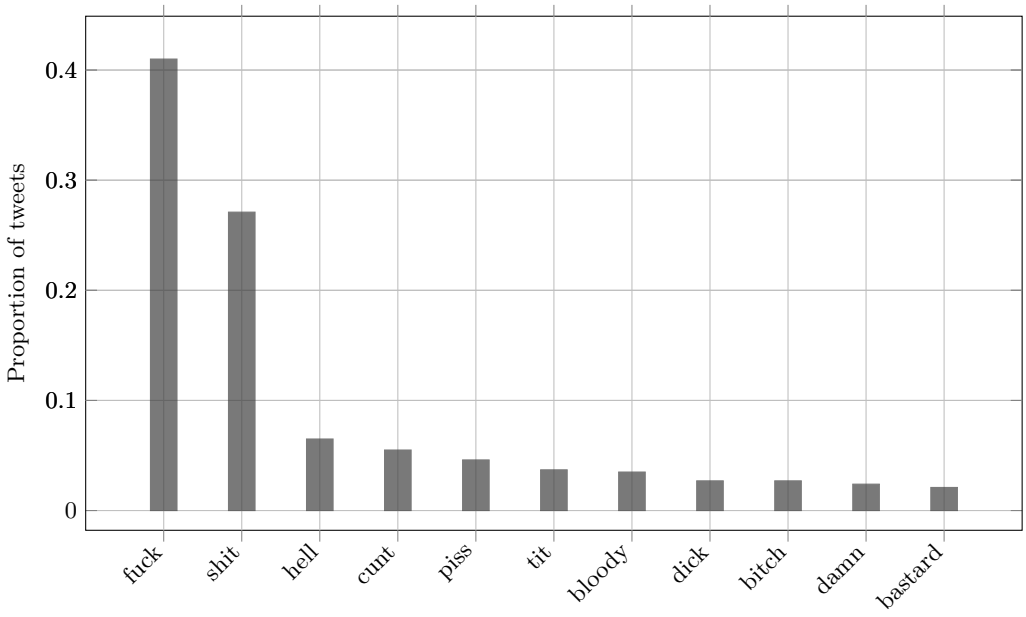


Fig. 3: Most common swear words found in swearing tweets published by male users.

4.4 Proportion of swearing tweets by gender per million words

Table 3 presents the proportions of use of all the swear words we took into account for both genders. As we mentioned before, there is an imbalance in the number of male and female users, as well as in the number of tweets for each gender. Thus, raw percentages would have been useless in that they are not comparable in such situations. To be able to efficiently

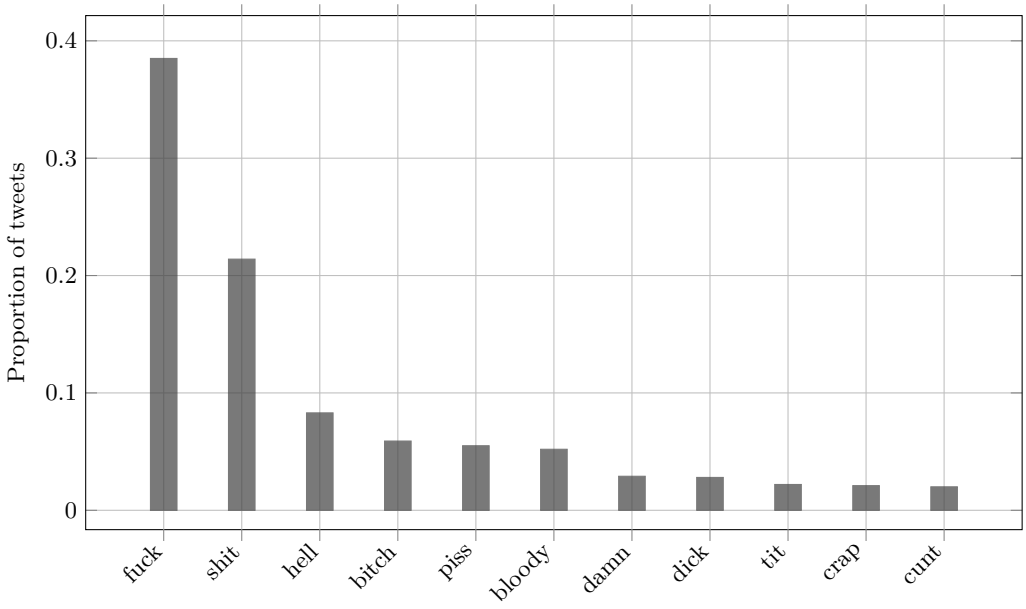


Fig. 4: Most common swear words found in swearing tweets published by female users.

compare the use of swear words by women and men, we calculated the number of instances of each swear word there would be in one million words. This then gives us an objective value on which to base our analyses. For each swear word, we calculated the log-likelihood (LL) score, which is based on the null hypothesis that there is no difference between the observed relative frequencies of a given swear word in the two corpora (i.e. female tweets and male tweets). We can reject the null hypothesis at the level of $p < 0.01$ when the LL value is greater than 6.63 (Rayson et al., 2004). In that case, we consider this word to be characteristic of men or of women. In Table 3, the three most statistically significant words for women and men are highlighted. These words are, in descending order of significance, *bitch*, *bloody* and *hell* for women, and *cunt*, *tit* and *fuck* for men. It would seem that some of the findings of McEnery (2006) are verified here, as in his study of the use of swear words of women and men on MySpace, he found that *fucking, fuck, jesus, cunt* and *fucker* were more typical of males, and *god, bloody, pig, hell, bugger, bitch, pissed, arsed, shit* and *piss* were more typical of females. However, although some of the most significant words for women and men in our sample were also significant for the MySpace users, the ranking of those words is different. *Cunt* is now the most significant word for men, and *bitch* the most significant for women, which may suggest an evolution in the gendered preferences of swear words. However, it may also be due to different ways of swearing and topical differences triggered by the two social media in question (MySpace and Twitter).

4.5 Average ratio of swearing tweets per day by gender and age

As mentioned earlier, Twitter's API enables us to collect many information along with tweets themselves. The time at which those tweets are published is one of them. Figures 5 and 6

Table 3: Frequency of swear words by gender per million words.

fuck	**1719.21**	**2105.07**	**Men**	**179.15**
shit	996.36	1198.84	Men	85.96
ass	63.48	63.48	Neither	0
bitch	**262.02**	**142.88**	**Women**	**168.1**
nigga	10.97	18.82	Men	9.51
hell	**344.57**	**309.89**	**Women**	**8.54**
whore	11.23	10.43	Neither	0.14
dick	124.34	133.64	Neither	1.54
piss	257.84	253.08	Neither	0.21
pussy	19.33	33.7	Men	17.92
slut	20.11	15.74	Neither	2.49
tit	**88.03**	**195.24**	**Men**	**187.47**
fag	18.02	30.28	Men	14.3
damn	117.81	102.84	Neither	4.73
cunt	**94.3**	**295.86**	**Men**	**487.84**
cum	8.62	17.79	Men	14.7
cock	82.02	93.08	Neither	3.22
retard	11.23	26.86	Men	29.71
blowjob	1.04	1.88	Neither	1.1
wanker	21.42	49.96	Men	52.82
bastard	68.44	110.54	Men	45.49
prick	47.02	84.36	Men	48.8
bollocks	14.62	37.3	Men	45.96
bugger	18.8	18.3	Neither	0.03
bloody	**227.01**	**173.51**	**Women**	**33.47**
crap	100.05	89.66	Neither	2.64

represent the average swearing ratio throughout the day for male and female users aged between 12-18 years old. The global patterns are the same for women and men for both age groups, the highest peak of swearing ratio being located in every case between 2am and 5am. In other words, this period is the one in which the proportion of swearing tweets compared to non-swearing tweets is the greatest.

During the day, the pattern seems to be the same for both genders from both age groups as well, since the swearing ratio for both genders keeps increasing throughout the day. As Wenbo et al. (2014) noted, we notice that the global pattern of swearing tweets corresponds to the standard activity of human life, as users start swearing between 6am and 7am, when people usually wake up, and gradually increases throughout the day. Interestingly, we observe

that there is a downfall in the swearing ratio around dinner time (around 7pm and 8pm), which suggests that people tweet, or swear less at that moment. However the ratio increases drastically after that period. As studies have shown, one of the main functions of swearing is to express strong emotions like anger, joy or sadness (Allan & Burridge, 2006; Jay & Janschewitz, 2008). Considering these interpretations, the fact that the ratio for men is constantly higher than women may suggest that they feel more comfortable expressing these kinds of emotions than women.

Thus, apart from the differences between genders, what these figures suggest is that both women and men have the same attitude regarding swearing throughout the day. Even if the way women and men swear can differ lexically or quantitatively (as shown in Table 3), some aspects of swear words usage are the same, and apparently the way women and men use swear words according to the time of the day is something which is common to both genders in our corpus.

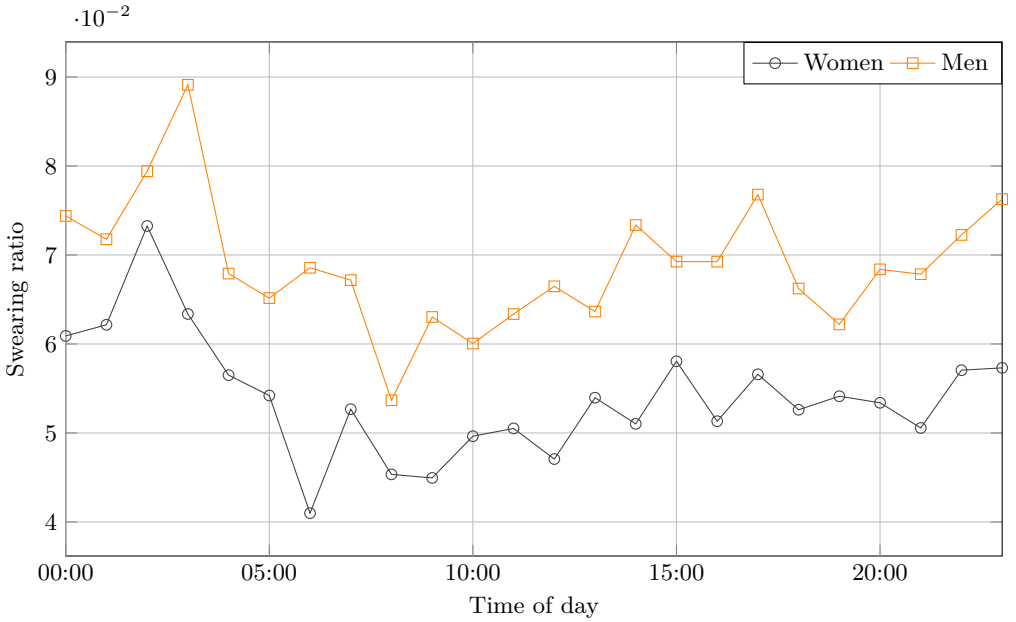


Fig. 5: Swearing ratio versus time of day in $\mathcal{C}^{f}_{12-18}$.

4.6 Named-entity recognition

Named-entity recognition (NER) enables us to automatically locate specific elements in tweets, more precisely the names of people, organizations or locations. To perform this task, we use a software (Finkel et al., 2005) which implements a NER method that relies on classification rules based on features of the word sequences that constitute tweets.

Table 4 reveals that on average, men use named entities more than women. Also, for both women and men, users tend to mention named entities consistently more as they get older.

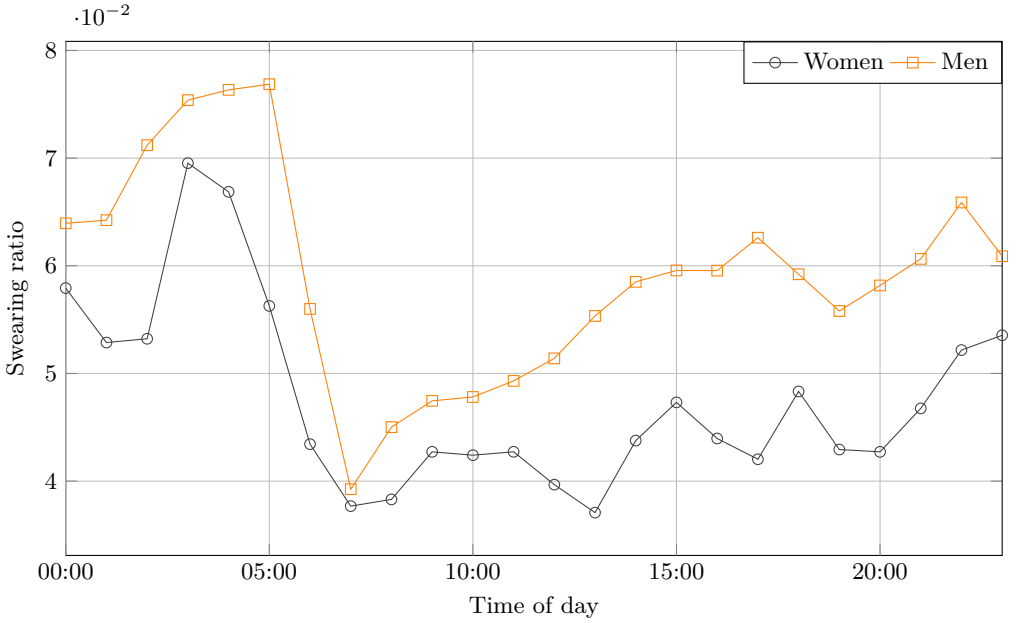


Fig. 6: Swearing ratio versus time of day in $\mathcal{C}^m_{12-18}$.

Figures 7 and 8 present detailed proportions of named entities per gender and age group in swearing tweets. This shows that whatever their age, both women and men majoritarily mention named entities referring to people when swearing. However, what differs is the fact that women from every age groups seem to favor locations over men, who prefer mentioning organizations. This method then highlights the fact that as far as swearing is concerned, context plays a big role. We suggest that these differences may point at gendered differences in the topics women and men focus on, at least when they swear, which may reveal the fact that the pragmatic functions of swear words for women and men of the same age groups may differ. However, more qualitative analyses would be necessary to be able to confirm this hypothesis.

Table 4: Proportion of tweets that contain named entities.

	[12; 18]	[19; 30]	[31; 45]	Average
Women	10.28%	13.41%	14.60%	12.76%
Men	14.25%	19.67%	20.59%	18.18%

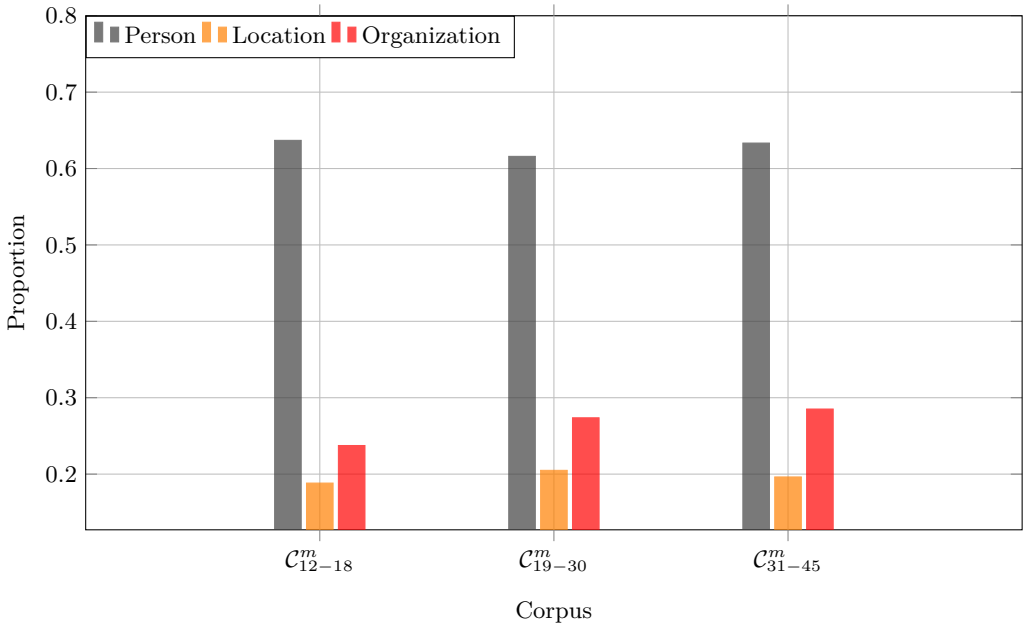


Fig. 7: Distribution of the types of named entities in swearing tweets published by men.

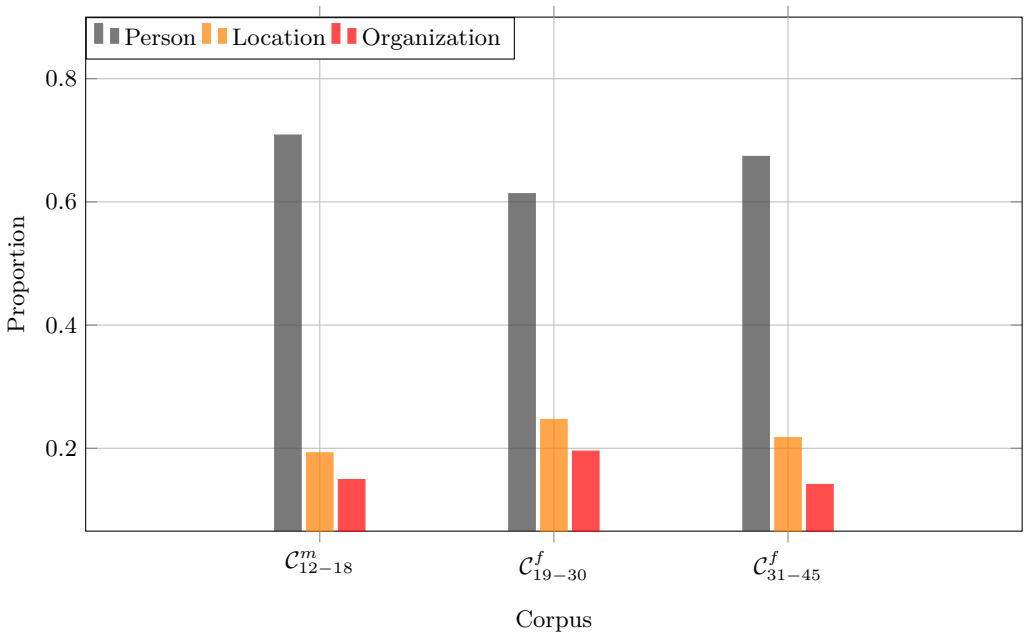


Fig. 8: Distribution of the types of named entities in swearing tweets published by women.

4.7 Event detection

In order to analyze the impact that real world events may have on Twitter discussions, we studied specific reactions on Twitter triggered by the most influential of these real world

events (e.g. the broadcast of a popular TV show, a political event, etc.) for users. We use Mention-Anomaly-Based Event Detection (MABED), a statistical method proposed by Guille & Favre (2015, 2014) for the detection of significant events from tweets. Thanks to this method, we are able to map both macro and micro levels of gendered reactions, as it describes each event it detects with a set of words, a time interval and a score that reflects the magnitude of impact of the event over users. Moreover, it is possible to analyze the tweets associated with these events, to understand their underlying composition, and the way swear words are used in our case for example.

Table 5 and Table 6 present as an example the ten most significant events detected by MABED for women and men aged 19-30. These events are numbered from 1 to 10 in decreasing order of significance. The 'event' column presents the keywords recognized as the most representative of the event, and the last column presents the percentage of tweets containing at least one swear word inside each event. Events marked as "spam" were events which were considered as such because one single user posted the same spam tweet very often, thus virtually generating keywords considered by MABED as relevant events. These spammers, though being a minority in our corpus as shown in Figure 1, create a considerable amount of noise for our event detection method and prevent a more accurate analysis of gendered events. It is however interesting to note that in this sample, spammers are twice as more present in the female corpus than in the male one, thus suggesting that spam accounts are more likely to adopt a female name.

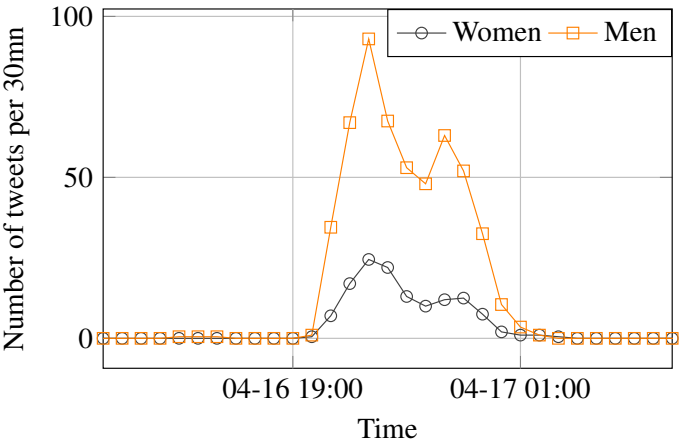


Fig. 9: Evolution of the number of tweets containing #bbcdebate.

Apart from spams, which have been manually identified as such, what the results from MABED reveal is that generally speaking, male events could be summarized by sports (boxing and soccer) and politics, and female events by media/entertainment (birth of the Royal baby, BGT (Britain's Got Talent)), sports (Grand National) and politics. Generally speaking, we observe that throughout those ten events, men use more swear words than women, and the event with the smallest number of swear words among men still contains

more of these than the event containing the greatest number of swear words among women. The events containing the most and least amount of swear words are highlighted for each gender, and this reveals that both women and men swear more when talking about politics.

What is interesting to notice is that on average, the proportion of swearing tweets in reaction to an event is higher (11.3% for men and 6.53% for women) than in classic interactions on Twitter (5.8% for men and 4.8%). It is also worth mentioning that the only common topic between women and men is the one containing the hashtag #bbcdebate. Figure 9 plots the evolution of the number of gendered tweets containing this hashtag. Though men tweet consistently more about that hashtag than women (but it must be reminded that this graph presents the raw number of tweets, so the gendered imbalance may be explained by the fact that we have more men than women in our corpus), we observe that the two patterns are very similar, and that the events are both detected roughly when the broadcast of the debate starts on television, and gradually decrease after the broadcast is over, as it triggers fewer and fewer reactions. The proportion of swearing tweets inside this common event does not differ much between women and men, which may again suggest that gendered differences in swearing are not triggered by gender alone, but by the context in which swearing occurs. In other words, women and men in the exact same context would not differ much in the linguistic attitudes they display. This would imply that swear words are not so much gendered as contextualized, which would correspond to other studies pointing to the fact that when considering gendered speech patterns, the context of use plays a greater role than gender alone (Eckert, 2008; Bamman et al., 2014; Baker, 2014; Holmes, 1995; Ladegaard, 2004). In our case, further qualitative research would however be needed to confirm or refute that hypothesis.

Table 5: Top 10 most impactful events detected from $\mathcal{C}^{m}_{19-30}$

Rank	Event	% of swearing tweets
1	*spam*	
2	fight maywheather he maypac watch pacquiao	13.00%
3	bbcdebate ed farage about natalie milliband	10.70%
4	messi ronaldo best goal ever boateng world lionel what	10.10%
5	snp seats have	9.60%
6	**tories labour have more you**	**15.50%**
7	*spam*	
8	bournemouth league premier next all well play football	11.70%
9	**exit polls ge2015 have wrong hope right lib**	**9.40%**
10	seat his lost	10.40%

Table 6: Top 10 most impactful events detected from $\mathcal{C}^{f}_{19-30}$.

Rank	Event	% of swearing tweets
1	*spam*	
2	**royalbaby princess kate girl charlotte diana baby name**	**2.50%**
3	*spam*	
4	bgt dog antanddec ant me max omg	6.60%
5	*spam*	
6	baby royalbaby girl princess kate	4.40%
7	National grandnational bets	7.90%
8	*spam*	
9	**bbcdebate nigel up ed nhs would**	**9.20%**
10	Grand national bets	8.60%

5 Limitations

This study presents certain limits. The first one concerns the way we categorized users according to their age. Though it has some advantages, it is not perfect, as some users will have children before age 30, or will leave school before age 18, so the linguistic patterns potentially influenced by those social phenomena may differ.

Another potential problem is that we did not include hashtags in our swear word detection methods, and hashtags often contain swear words, thus potentially limiting our data in this regard.

A manual verification of the information provided in the description of a lot of users in our sample reveals that many are students. Even if it sounds normal, as the most represented age group is the 19-30, there may thus exist a bias towards this category of users.

6 Conclusion

In this article, we tried to give hints about new methods which could be used to analyze specific sociolinguistic parameters on Twitter. For that purpose, we analyzed the data of a corpus of about one million tweets from users for whom we could infer both the age and the gender. Even if our data would need to be analyzed more thoroughly and qualitatively in order to draw more generalizable conclusions, our goal here was to show that by combining techniques from both computer science and linguistics, it is possible to provide innovative ways of studying the way women and men swear on Twitter. These tools showed that beyond mere quantitative data which could lead to erroneous impressions and generalizations on the reasons why women and men swear, contextual parameters are sometimes more important in being able to determine what is influential, as we concluded with the event detection and NER analyses. This work is then the continuation of prior studies which showed that gender is often enacted in subtil ways, hence the necessity to develop more tools to explore these

questions. We believe that some of the tools presented here can be used improved in future research based on Twitter data, so that the analyses presented in this paper can be refined, especially in order to be more qualitative.

Bibliography

[1] Allan, K., & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language.* Cambridge University Press.

[2] Baker, P. (2014). *Using corpora to analyze gender.* Bloomsbury Publishing.

[3] Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 135–160.

[4] Baruch, Y., & Jenkins, S. (2007). Swearing at work and permissive leadership culture: When anti-social becomes social and incivility is acceptable. *Leadership & Organiation Development Journal*, 28(6), 492–507.

[5] Beers Fägersten, K. (2012). *Who's swearing now? The social aspects of conversational swearing.* Cambridge Scholars Publishing.

[6] Coates, J. (2004). *Women, men and language: a sociolinguistic account of gender differences in language.* Pearson.

[7] Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12, 453–476.

[8] Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd annual meeting of the association for computational linguistics* (pp. 363–370).

[9] Guille, A., & Favre, C. (2014). Mention-anomaly-based event detection and tracking in Twitter. In *Proceedings of the IEEE/ACM international conference on advances in social network analysis and mining* (p. 375-382).

[10] Guille, A., & Favre, C. (2015). Event detection, tracking and visualization in Twitter: a mention anomaly based approach. *Social Network Analysis and Mining*, 5(1), 1-18.

[11] Hammons, J. W. (2012). *Wgaf: Swearing, social structure and solidarity in an online community* (Unpublished doctoral dissertation). Ball State University, Indiana.

[12] Harris, R. (1990). Lars porsena revisited. In *The state of the language* (p. 411-421). Faber.

[13] Holmes, J. (1995). *Women, men and politeness.*

[14] Hughes, G. I. (2006). *An encyclopedia of swearing: The social history of oaths, profanity, foul language, and ethnic slurs in the English-speaking world.* Sharpe.

[15] Hughes, S. E. (1992). Expletives of lower working-class women. *Language in Society,* 21(2), 291-303.

[16] Jay, T. (1992). *Cursing in America: a psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards, and on the streets.* John Benjamins.

[17] Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture,* 4(2), 267–288.

[18] Ladegaard, H. J. (2004). Politeness in young children's speech: context, peer group influence and pragmatic competence. *Journal of Pragmatics*, 36.

[19] Lakoff, R. (2004). *Language and woman's place.* Harper and Row.

[20] McEnery, T. (2006). *Swearing in English.* Routledge.

[21] Mercury, R. E. (1995). Swearing: A “bad” part of language; a good part of language learning. *TESL Canada Journal*, 13(1), 268–36.

[22] Murray, T. E. (2012). Swearing as a function of gender in the language of Midwestern American college students. In *A cultural approach to interpersonal communication: Essential readings* (pp. 233–241). Blackwell.

[23] Rayson, P., Berridge, D., & Francis, B. (2004). Extending the cochran rule for the comparison of word frequencies between corpora. In Proceedings of the 7th international conference on statistical analysis of textual data (p. 926-936).

[24] Smith, A., & Brewer, J. (2012). *Twitter use 2012* (Tech. Rep.). Pew Research Center.

[25] Stapleton, K. (2003). Gender and swearing: A community practice. *Women's and Language*, 26(2).

[26] Stapleton, K. (2010). Swearing. *Interpersonal pragmatics*, 289–305.

[27] Thelwall, M. (2008). Fk yea i swear: Cursing and gender in a corpus of MySpace pages. *Corpora*, 3(1), 83–107.

[28] Wenbo, W., Lu, C., Krishnaprasad, T., & Amit, P. S. (2014). Cursing in English on Twitter. In *Proceedings of the ACM conference on computer supported cooperative work and social computing* (p. 415-425).

Mapping the structure of the global maker laboratories community through Twitter connections

Massimo Menichinelli

Aalto University
School of Art, Design and Architecture
Department of Media
Media Lab Helsinki
P.O.Box 31000,00076 Aalto, Finland.
massimo.menichinelli@aalto.fi

Abstract. The Maker Movement is currently considered an interesting and promising phenomenon with social and economic implications, especially through a series of locally implemented but globally connected laboratories for making. The social structure of such global community as a whole has not been examined yet. This research proposes therefore to analyze the connections of following and being followed among the Twitter accounts of such Maker laboratories, as a proxy of their pattern of trust and influence. The Twitter accounts were manually gathered in Twitter lists with the help of the search functionalities of Twitter, and the data was later accessed and reconstructed with a script based on the NetworkX library. The network was reconstructed from the ego-networks of each account on the examined lists, and accounts which were not part of the lists were removed. The resulting network shows a relevant polarity between roughly Fab Labs on one side and Hackerspaces, Makerspaces and TechShops on the others. Further analysis of the data revealed that the Fab Lab sub-community is much more diversified and less homogeneous than the other part, with further insights on the impact of the organization of such laboratories and their networks. Furthermore, the analysis of some centrality measures showed how few nodes bridge these polarities, while most of the nodes are very close to each other and the Fab Lab network has a more relevant and distributed influence. A possible direction for improving this research along multiple dimensions is proposed as a conclusion.

Keywords: Fab Lab, Makerspace, Hackerspace, TechShop, Makers, Social Network Analysis, Twitter, Community

1 Introduction

The Maker Movement is currently considered an interesting and promising phenomenon with social and economic implications (Anderson, 2012; Hatch, 2014). The "Maker" term is quite broad, and there are currently many discussions regarding how to clearly identify Makers. Generally, however, Makers are usually defined as a people who are interested in designing, making, repairing physical objects with digital tools, and that have the required skills for such activities and share their projects, experiences or time in local laboratories.

The emergence of Makers has been facilitated by three factors: (1) the growing accessibility of tools for manufacturing physical objects; (2) the evolution of web platforms from just discussion to even design, prototyping and manufacturing; (3) the growing number of local laboratories for individual and collaborative design and making. On one side, the development of manufacturing technologies and especially technologies based on digital fabrication (Gershenfeld, 2005, 2012) has been increasing the quality of manufacturing processes, it has been lowering the cost of such technologies and processes which consequently started to be more widespread available. Many open source software and hardware projects like the RepRap 3D Printer (Sells, Bailard, Smith, Bowyer, & Olliver, 2009) or the Arduino development board (Banzi, 2009) lowered the barriers to creating physical objects regarding the low price, the local availability, the easiness of building and repairing and the easiness of studying and developing projects on top of them. On another side, the emergence of the Web 2.0 or Social Web phenomenon (O'Reilly, 2005) generated multiple experiments of online platform for the emergence of collaborative communities, both non-profit and for-profit, that represented a further development of the first online communities, bringing them to online discussion to online software, hardware and design development, prototyping and manufacturing with platforms like GitHub[1], Thingiverse[2], UpVerter[3], Shapeways[4] and Ponoko[5] (Anderson, 2012; Benkler, 2002; Howe, 2006; Tapscott & Williams, 2006). On a third side, the emergence of many different formats of Maker facilities like Hackerspaces, Makerspaces, TechShops and Fab Labs made these technologies and processes locally available in a growing number of localities (Anderson, 2012; Cavalcanti, 2013; Gershenfeld, 2005; Hatch, 2014; Maxigas, 2012; Tweney, 2009). The interactions among these three phenomena are creating the conditions for the emergence of the Open and Distributed Manufacturing scenario, where design and production are re-localized, supply chains redesigned and new business generated, all based on local communities which are connected globally (Bauwens, 2009).

Therefore, it is especially through a series of such local but globally distributed laboratories, that this phenomenon of the Maker Movement (i.e. the people) and of Open and Distributed Manufacturing (i.e. the technologies and the business) is taking place and growing with new members, technologies and businesses. Such laboratories have different names and formats like Hackerspaces, Makerspaces, Fab Labs, TechShops and so on. Each format has a different story, culture, typical business plan and tools for networking among them. Boundaries among these formats are not always well defined, as there are both places that may fit into different formats and Makers who participates in different formats. As a whole, these laboratories constitute the local dimension of the whole Make Movement. Currently, it is widely considered that such laboratories (and the whole Maker Movement) constitute a global community thanks to the use of ICT technologies for communication and coordination. The social dimension of the local community of some specific laboratory has been investigated, but

[1] `https://github.com/`
[2] `http://www.thingiverse.com/`
[3] `https://upverter.com/`
[4] `http://www.shapeways.com/`
[5] `https://www.ponoko.com/`

the global community as whole has not been examined yet. For example, some researchers analyzed the local community that gathers around Fab Lab Amsterdam (Ghalim, 2013; Maldini, 2014); others focused on the national community of Fab Labs, Hackerspaces and Makerspaces in Italy (Menichinelli & Ranellucci, 2015) or the national community of Makers in Italy (Bianchini, Menichinelli, Maffei, Bombardi, & Carosi, 2015). Thus there is no evidence yet that such a movement really constitutes a global community or just single laboratories or patterns of single laboratories and of more coordinated and connected laboratories.

The aim of this research is to shed a first light on the social dimension of the global phenomenon of Maker laboratories. The task of examining such a wide phenomenon can be extremely difficult because of its global and distributed nature, and because each format uses different tools and governance strategies for networking. This research therefore propose to analyze the Twitter accounts of such Maker laboratories, as a proxy of the global phenomenon. Twitter was not designed as a place for community building, but some investigations proved that it could actually give place to communities (Gruzd, Wellman, & Takhteyev, 2011). The goal of this research is of analyzing the patterns of connections among such Twitter accounts, in order to find which kind of social structures they have and which kind of community or sub-communities they form. In order to discover such structures, this research analyzes the connection among the Twitter accounts of such Maker laboratories and their connections of following / being followed, as a proxy of the disposition to communicate with other laboratories and of therefore the influence of laboratories in being listened to. Since such laboratories represents experiments in forms of networked and distributed organizations, such analysis could shed a first light on the role and distribution of influence and organization in such distributed systems.

2 Method

This research focused on analyzing the connections of following or being followed among the Twitter accounts of such Maker laboratories as a first step in analyzing such global community. The Twitter accounts were manually gathered with the help of the search functionalities of Twitter, in order for the researcher to be able to select those account that were really related to the context of the research: several Twitter accounts have a name that could suggest their belonging to the global community, but upon a closer examination they showed to just use the same keywords. The accounts that were considered were of laboratories, events or organizations related to the global Maker community, since these entities are strongly linked. The accounts were initially separated in three Twitter lists, but these lists were joined later during the analysis of the data, therefore constituting one single list.

The social network of following or being followed connections among the accounts was built with an ego-network of each account: for each account analyzed, Twitter accounts that followed or were followed by it were gathered. Before analyzing the data, all the Twitter accounts that did not belong to the lists were removed. In this way, only the Twitter accounts of the global Maker laboratories were finally analyzed. The data was downloaded and reconstructed into a graph with a custom Python script, based on the NetworkX library (Hagberg, Schult, & Swart, 2008) that accesses the lists trough the Twitter API (Menichinelli,

2015). The script has been released with a Free Software / Open Source license and its code and development can be accessed, commented and contributed to online on GitHub[6]. The obtained data was later analyzed and visualized with Gephi (Bastian, Heymann, & Jacomy, 2009) with the a Force-Directed layout (Jacomy, Venturini, Heymann & Bastian, 2014) and NetworkX with the Matplotlib and Seaborn Python libraries (Hunter, 2007; Michael Waskom et al., 2014). The resulting network was first analyzed in terms of the different sub-communities, in order to understand its structure, and then in terms of several centrality measures, in order to understand the distribution of influence among the nodes. Before exporting the final data, the Twitter accounts were anonymized; however the awareness of the identity of the Twitter accounts proved to be useful during the analysis, in order to understand the sub-communities present in the global network. The research can be replicated with the same software developed for it (Menichinelli, 2015) and one or more Twitter lists[7].

3 Results

3.1 General information

The development of the script and the analysis of the data took place along several months during 2014 and 2015; one intermediary analysis was done on the 8th of January 2015 and the final analysis was done on the 26th of May 2015. The act of following or being followed on Twitter is asymmetric, therefore the collected graph was reconstructed as a directed graph. The full network of accounts who were followed or followed the Twitter accounts of the Maker laboratories consisted of 499,681 nodes and 1,098,373 edges: there were errors with accessing the information of only 10 Twitter accounts, who where most likely inaccessible as a consequence of a privacy settings. The final network consisting only of Twitter accounts of Maker laboratories consists of 946 nodes and 29,821 edges. The density of the network is quite low at 0.033: a common value for large networks. The average path length is 2.755, indicating that nodes are highly connected and information can flow quickly among them. The diameter is 10, indicating that there are 10 connections between the farthest nodes in the network: this is a small path, if we consider that the network has 946 nodes. These values could be considered a first measurement of the size of the global Maker laboratories: it is a difficult task to calculate the number of laboratories since there are several online platforms for mapping them. Furthermore, since most of these laboratories start as bottom-up initiatives, it is difficult for such platforms to keep track of the development of the global Maker laboratories community.

3.2 Modularity measure: the structure of the sub-communities

The most relevant feature of the network is how it is subdivided into two main polarities, which are only loosely connected by few nodes. These two polarities consists roughly of

[6] `https://github.com/openp2pdesign/Twitter-Lists-SNA-EgoNetworks`
[7] The raw anonymized data can be downloaded here: `http://dfn.link/ktq2`

Fab Labs on one side and mainly of Hackerspaces, Makerspaces and TechShops on the other side. In order to understand the different sub-communities that constitute such global phenomenon, the network was analyzed with a modularity detection. The measurement was done in in Gephi (Bastian et al., 2009), which uses as specific algorithm for modularity detection (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) with a resolution parameter for identifying more or less communities (Lambiotte, Delvenne, & Barahona, 2008). At first, the network was analyzed with a resolution of 1.0, and then with a resolution of 0.5, going then deeper into its sub-communities.

In the first case (Table 1), with a resolution of 1.0, 41 communities were found but only three big communities are relevant for their size: Hackerspaces, Makerspaces and TechShops together on the right, and Fab Labs on the left and a separated subset of French Fab Labs on the lower part of the left, pointing out a first subdivision of the Fab Lab community and instead a more homogeneous structure of Hackerspaces, Makerspaces and TechShops (Figure 1). As a whole, Makerspaces and Hackerspaces constitute 53.28% of the network, while Fab Labs are 42.07% of the network.

Table 1: Main sub-communities identified with modularity detection with a resolution of 1.0

Size of the sub-community related to the global network	Sub-community identified
53.28%	Makerspaces and hackerspaces
37.1%	Fab Labs
4.97%	A subset of French Fab Labs
$< 1.0\%$	Several smaller sub-communities

In the second case (Table 2), with a resolution of 0.5, the number of sub-communities rises to 54: on the right, we can clearly distinguish Hackerspaces (in green), Makerspaces (in yellow) and TechShops (in dark red) (Figure 2). Counting with a longer history, Hackerspaces constitute 32.66% of the network, while Makerspaces constitute 16.7% of the network. TechShops, which is the format that mostly related to a franchising of Maker laboratories, constitutes only 1.48% of the network. This is consistent with the common idea that these are connected but somehow different format of laboratories: at this resolution therefore we can identify some common formats. On the other side of the network on the left, at the same resolution, we find many more sub-communities, pointing out how the Fab Lab network is a much more diversified and articulated network of laboratories: a union of sub-communities more than a homogeneous community. While at this resolution we find the main Hackerspaces, Makerspaces and TechShops sub-communities, on the Fab Lab side there is a much more diversified structure. Few sub-communities on the Fab Lab side are related to a country (French, Italy, Spain), others are more mixed, but with a more prominent country (Netherlands, Japan). Finally, more sub-communities are constituted by laboratories from mixed countries.

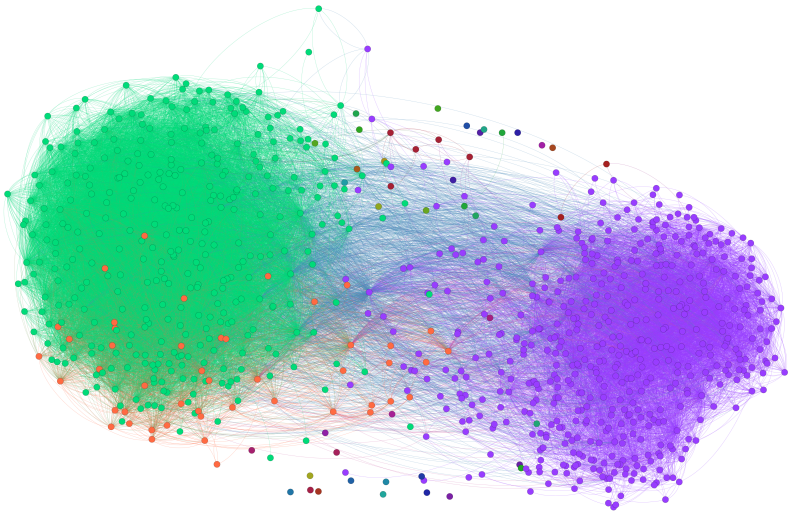


Fig. 1: The main sub-communities found with a resolution of 1.0 (highlighted with a different color).

Table 2: Main sub-communities identified with modularity detection with a resolution of 0.5

Size of the sub-community related to the global network	Sub-community identified
32.66%	Hackerspaces
16.7%	Makerspaces
6.66%	A subset of French Fab Labs, Hackerspaces, Makerspaces
6.13%	Most of Italian Fab Labs
5.71%	Mixed Fab Labs from all over the world
5.6%	Mixed Fab Labs from all over the world
3.7%	Most of Spanish Fab Labs
3.7%	Mixed Fab Labs from all over the world
3.38%	Mixed Fab Labs from all over the world
2.11%	Mixed Fab Labs from all over the world
2.11%	Mixed Fab Labs from all over the world, with especially Dutch laboratories
1.59%	Mixed Fab Labs from all over the world, with especially Japanese laboratories
1.59%	Mixed Fab Labs from all over the world
1.48%	TechShops
1.37%	Mixed Fab Labs from all over the world
$< 1.0\%$	Several smaller sub-communities

3.3 Centrality measures: importance and influence in the network

After having identified the structure of the sub-communities in the whole network, several centrality measurements were calculated in order to understand the health and the distribution

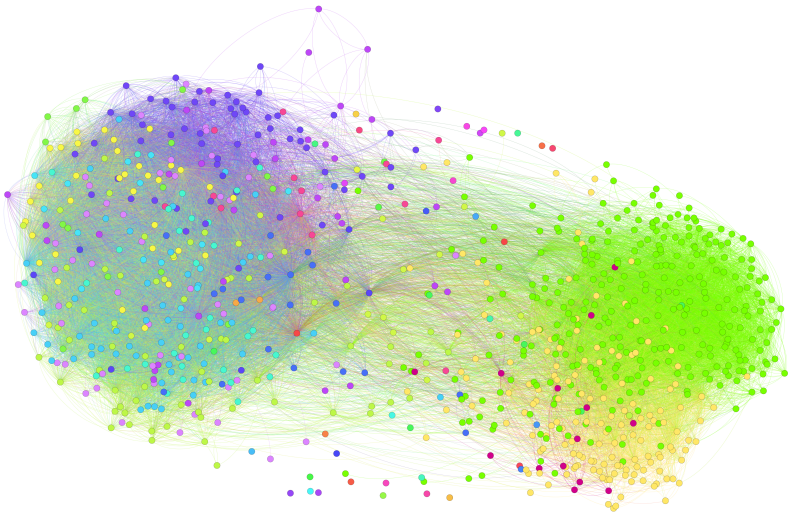


Fig. 2: The sub-communities found with a resolution of 0.5 (highlighted with a different color).

of influence in all the network and the differences in the sub-communities. Degree centrality was calculated first: the number of incoming and outgoing connections for each nodes is a first measurement of the importance of each node. Since the analyzed network is a directed graph, in-degree and out-degree were calculated: however the results are quite similar, with only few nodes more prominent on out-degree than in-degree. The average degree is 31.523, with only few nodes have an outstanding degree: the network is quite homogeneous beside these exceptions, and the distribution of degree follows a common power-law distribution (Figure 3). However, such an average degree is quite high, meaning that on average each laboratory is connected to 31 other laboratories. A higher number of nodes with higher degree values can be found in the Fab Lab sub-community compared to the Hackerspaces, Makerspaces and TechShops sub-community (Figure 4). The nodes with a higher degree can be found in the Fab Lab sub-community but at the interface with the sub-community of Hackerspaces, Makerspaces and TechShops.

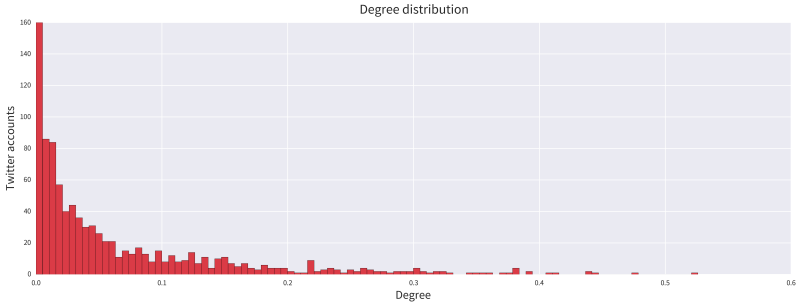


Fig. 3: Distribution of Degree centrality among the Twitter accounts gathered.

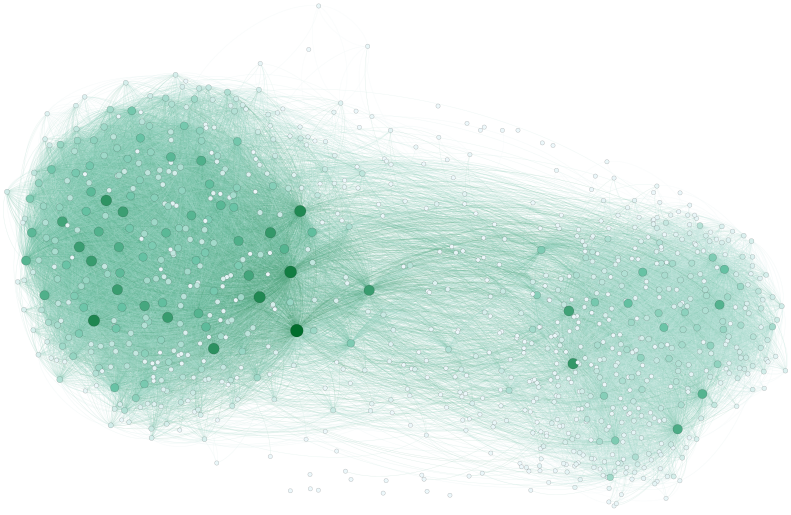


Fig. 4: Distribution of degree centrality in the network (highlighted with a color and size related to the value).

Betweenness centrality was then measured (Brandes, 2001) in order to understand the nodes who acts as a bridge among the many sub-communities found. Another common power-law distribution was found (Figure 5), but this time with a much higher number of nodes with a very low or close to zero value. Betweenness centrality is less uniformly distributed than degree centrality, with only few nodes in the Fab Lab sub-community (the same nodes noted before at the interface with the other sub-community) and then, with a lower value, few nodes in the Hackerspaces, Makerspaces and TechShops sub-community (Figure 6). Very few nodes act as a bridge among the two main sub-communities: 3-4 nodes on each sub-community interface with the other sub-community. Both main polarities are then connected by very few nodes.

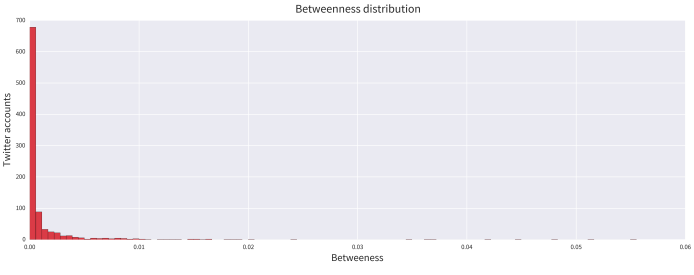


Fig. 5: Distribution of Betweenness centrality among the Twitter accounts gathered.

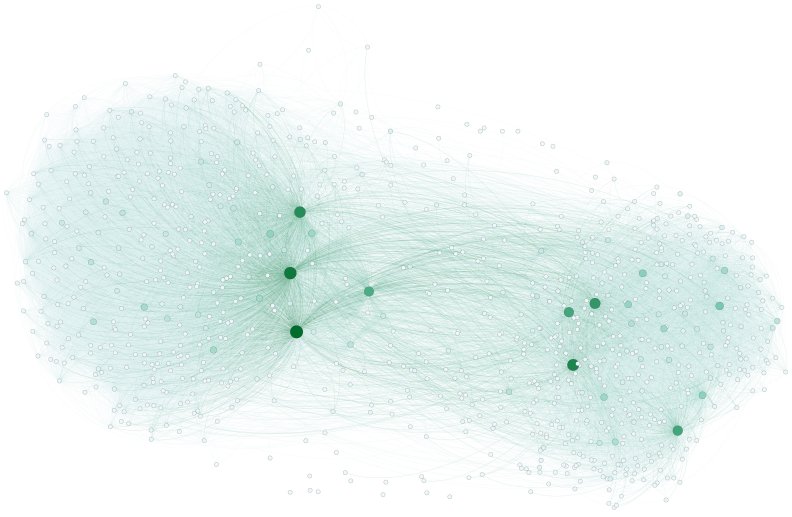


Fig. 6: Distribution of Betweenness centrality in the network (highlighted with a color and size related to the value).

In order to understand how close the nodes are compared to all the other nodes in the network, closeness centrality was calculated (Freeman, 1978). Here instead we have a different distribution: there is a peak at zero, and then the distribution is a normal one (Figure 7). Therefore, except for several nodes (21.67% of the whole network) that are far from the other nodes in the network (as they are disconnected regarding in-degree or out-degree), most of the network (78.32% of the whole network) has a more uniform distribution of closeness, showcasing short distances among nodes and therefore a common ability to spread information quickly (Figure 8). Both main polarities are strongly connected, where all nodes can easily communicate with each other and information can spread quickly: most of the laboratories are uniformly closed to each other, as in a real global community.

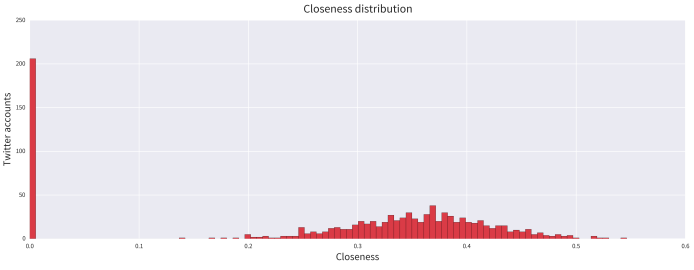


Fig. 7: Distribution of Closeness centrality among the Twitter accounts gathered.

Fig. 8: Distribution of Closeness centrality in the network (highlighted with a color and size related to the value).

In order to measure the importance of each node related to its neighbors, Eigenvector centrality and PageRank centrality (Brin & Page, 1998; Page, Brin, Motwani, & Winograd, 1999) (with a probability of 0.85) were calculated. These values measure the influence of each node on its neighbors and the network. Both measurements presents a common power-law distribution (Figure 9-Figure 11). There are differences, however, in the distribution of the measured values in the main sub-communities. Regarding Eigenvector centrality, nodes get more importance according to the importance of the neighbors: in this case the Fab Lab sub-community has clearly almost all the higher values, compared to the other part of the network (Figure 10). Nodes in the Fab Lab sub-community are clearly more influential in the whole Maker laboratories global community than the other laboratories, and this influence takes place with a power-law distribution where nodes with an higher influence are concentrated in the Fab Lab part of the network. Regarding PageRank centrality, it represents the likelihood that a Twitter account will reach a particular Twitter account in the network following the connections among all of the accounts. In this case, the value measured shows a power-law distribution as well but is concentrated only in very few nodes instead, especially in the Hackerspaces, Makerspaces and TechShops sub-community, with few nodes with a lower value but clearly identifiable in the Fab Lab sub-community (Figure 12).

4 Conclusions

The data obtained with this research give first insights about the structure and the distribution of influence and trust in the global Maker laboratories community on Twitter. The sub-

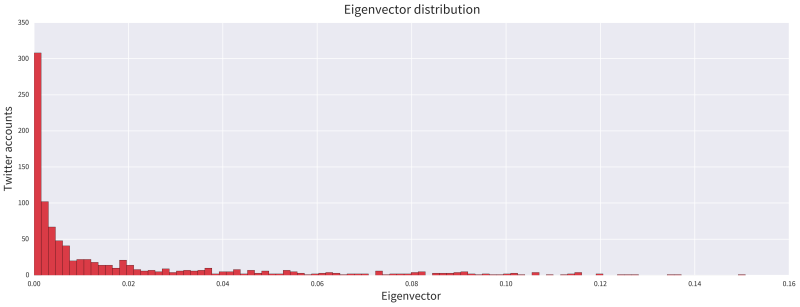


Fig. 9: Distribution of Eigenvector centrality among the Twitter accounts gathered.

Fig. 10: Distribution of Eigenvector centrality in the network (highlighted with a color and size related to the value).

Fig. 11: Distribution of PageRank centrality among the Twitter accounts gathered.

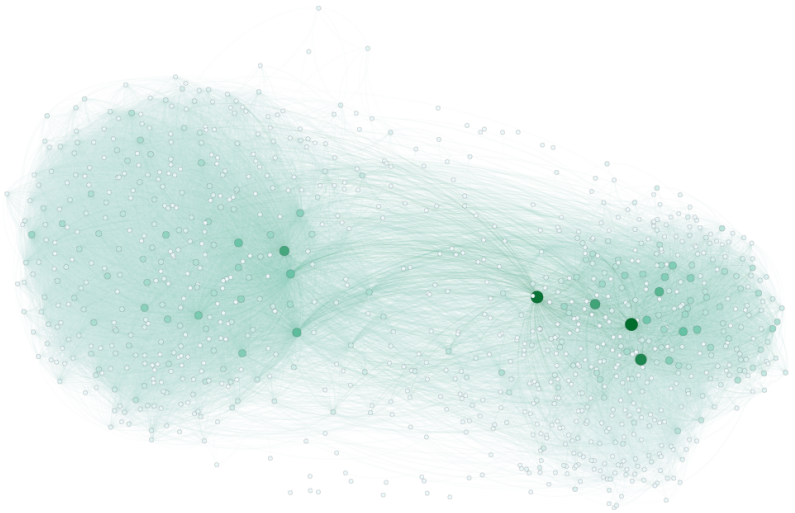


Fig. 12: Distribution of PageRank centrality in the network (highlighted with a color and size related to the value).

communities obtained give insights about the social structure of the global Maker laboratories community, showing how Hackerspaces, Makerspaces, and TechShops are more homogeneous groups compared to Fab Labs, and how these two polarities are well separated, even if with few connecting nodes. Some sub-communities are geographically located, but more are a mix of laboratories from different places, showing the global nature of the community. The presence of more sub-communities in the Fab Lab part could represent the presence of more groups inside the global community that could grow apart in the future, or it could represent how the global format and community of Fab Labs is growing to a more local level or how it could bridge distant localities together. Except few nodes with high degree and betweenness with a common power-law distribution, closeness is more uniformly distributed for connected nodes. Power-law distributions take place also regarding influence and trust: while Eigenvector centrality is much more concentrated in several nodes in the Fab Lab community, and PageRank centrality is concentrated in very few nodes among Makerspaces and Hackerspaces. Therefore, trust and influence take a different meaning in Fab Labs and in Makerspaces and Hackerspaces, according to how they are measured. The whole network is therefore constituted by several identifiable sub-communities that are organized in two polarities, with few nodes bridging them with a stronger role and trust; however most of the nodes are very close to each other in a strong community. Since each of the format of Maker laboratories follows different design processes, organizational structures and policies, these data could be used in order to further study the impact of such elements on the social structure of a global community, in order foster cohesion, communication and coordination.

This research represent a first analysis of the global Maker laboratories community on Twitter, as a proxy of the global connections among the laboratories. It focused only on the

connections of following or being followed by Maker laboratories Twitter accounts. Therefore, no real interactions were analyzed in this research: future research should investigate this global phenomenon along more multiple directions (Gruzd et al., 2011) in order to understand if such patterns of connections constitutes a real community and not just a consequence of activity or interest in Twitter. For example, further research could focus on the online discussion among Maker laboratories accounts, and also among them and users, and only among users, in order to understand the difference structure of the community as experienced by laboratories and as experienced by users.

Bibliography

[1] Anderson, C. (2012). *Makers: The New Industrial Revolution*. Crown Business.

[2] Banzi, M. (2009). Getting Started with Arduino (1st edition). Sebastopol: O'Reilly Media.

[3] Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In International AAAI Conference on Weblogs and Social Media (Vol. 2). Retrieved from `http://gephi.org/publications/gephi-bastian-feb09.pdf`

[4] Bauwens, M. (2009). The Emergence of Open Design and Open Manufacturing. We_ magazine, 02. Retrieved from `http://www.we-magazine.net/we-volume-02/`

[5] Benkler, Y. (2002). Coase's Penguin, or, Linux and The Nature of the Firm. The Yale Law Journal, 112. Retrieved from `http://www.yalelawjournal.org/the-yale-law-journal/content-pages/coase\%27s-penguin,-or,-linux-and-the-nature-of-the-firm/`

[6] Bianchini, M., Menichinelli, M., Maffei, S., Bombardi, F., & Carosi, A. (2015). Makers' Inquiry. Un'indagine socioeconomica sui makers italiani e su Make in Italy. Milano: Libraccio Editore. Retrieved from `http://makersinquiry.org/`

[7] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. `http://doi.org/10.1088/1742-5468/2008/10/P10008`

[8] Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology, 25, 163–177.

[9] Brin, S., & Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. In Proceedings of the Seventh International Conference on World Wide Web 7 (pp. 107–117). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V. Retrieved from `http://dl.acm.org/citation.cfm?id=297805.297827`

[10] Cavalcanti, G. (2013, May 22). Is it a Hackerspace, Makerspace, TechShop, or FabLab? Retrieved May 30, 2015, from `http://makezine.com/2013/05/22/the-difference-between-hackerspaces-makerspaces-techshops-and-fablabs/`

[11] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. Social Networks, 215.

[12] Gershenfeld, N. (2005). FAB: The Coming Revolution on Your Desktop–From Personal Computers to Personal Fabrication. Basic Books.

[13] Gershenfeld, N. (2012). How to Make Almost Anything: The Digital Fabrication Revolution. Foreign Affairs, 91, 43–57.

[14] Ghalim, A. (2013). Fabbing Practices: An Ethnography in Fab Lab Amsterdam (Master's Thesis). Universiteit van Amsterdam (New Media and Culture Studies), Amsterdam. Retrieved from `http://www.scribd.com/doc/127598717/FABBING-PRACTICES-AN-ETHNOGRAPHY-IN-FAB-LAB-AMSTERDAM`

[15] Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. American Behavioral Scientist, 55(10), 1294–1318. http://doi.org/10.1177/0002764211409378

[16] Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008) (pp. 11–15). Pasadena, CA USA. Retrieved from http://math.lanl.gov/~hagberg/Publications/hagberg-2008-exploring.shtml

[17] Hatch, M. (2014). The maker movement manifesto. Rules for innovation in the new world of crafters, hackers, and tinkerers. New York: McGraw-Hill Education.

[18] Howe, J. (2006, June). The Rise of Crowdsourcing. Wired, 14(6). Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

[19] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing In Science & Engineering, 9(3), 90–95.

[20] Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9(6): e98679. doi: 10.1371/journal.pone.0098679.

[21] Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2008). Laplacian Dynamics and Multiscale Modular Structure in Networks. arXiv:0812.1770 [physics]. Retrieved from http://arxiv.org/abs/0812.1770

[22] Maldini, I. (2014). Digital makers: an ethnographic study of the FabLab Amsterdam users. In A Matter of Design. Making Society through Science and Technology. Retrieved from http://www.stsitalia.org/conferences/ocs/index.php/STSIC/AMD/paper/view/58

[23] Maxigas. (2012). Hacklabs and Hackerspaces. Tracing Two Genealogies. Journal of Peer Production, (2). Retrieved from http://peerproduction.net/issues/issue-2/

[24] Menichinelli, M. (2015). Twitter-Lists-SNA-EgoNetworks: v0.2 stable. http://doi.org/10.5281/zenodo.18209

[25] Menichinelli, M., & Ranellucci, A. (2015). Censimento dei Laboratori di Fabbricazione Digitale in Italia 2014. Roma: Fondazione Make in Italy CDB. Retrieved from http://www.makeinitaly.foundation/censimento-dei-laboratori-fabbricazione-digitale-in-italia/

[26] Michael Waskom, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, & Dan Allan. (2014). seaborn: v0.5.0 (November 2014). http://doi.org/10.5281/zenodo.12710

[27] O'Reilly, T. (2005, September 30). What Is Web 2.0 - O'Reilly Media. Retrieved March 31, 2012, from http://oreilly.com/web2/archive/what-is-web-20.html

[28] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Retrieved from http://ilpubs.stanford.edu:8090/422

[29] Sells, E., Bailard, S., Smith, Z., Bowyer, A., & Olliver, V. (2009). RepRap: The Replicating Rapid Prototyper: Maximizing Customizability by Breeding the Means of Production. In F. T. Piller & M. M. Tseng, Handbook of Research in Mass Customization and Personalization (pp. 568–580). World Scientific Publishing Company. Retrieved from http://www.worldscientific.com/doi/abs/10.1142/9789814280280_0028

[30] Tapscott, D., & Williams, A. D. (2006). Wikinomics: How Mass Collaboration Changes Everything. Portfolio Hardcover.

[31] Tweney, D. (2009, March 29). DIY Freaks Flock to "Hacker Spaces" Worldwide. Retrieved March 31, 2012, from http://www.wired.com/gadgetlab/2009/03/hackerspaces/

MapTwitter tribal language(s)

Nelleke Oostdijk and Hans van Halteren

CLS
Radboud University
Nijmegen, Netherlands
{n.oostdijk;hvh}@let.ru.nl

Abstract. The highly variable nature of Twitter langu age use has been amply noted in a wide variety of studies, where quite often it is found to present a problem to existing tools used for the automatic processing of tweets. In this paper we investigate the nature of the variability, with a focus on the syntactic constructions that are used. In a first exploratory study we cluster hashtags in Dutch tweets on the basis of the users contributing to those hashtags. On further investigation of four larger coherent groups of hashtags (related to school, employment, politics, and appreciation), we find that there are clear differences in the overall language use in the various groups, also at the syntactic level. We discuss the clustering, specific language use patterns for the hashtag groups, and the implications for the automatic processing of tweets for research.

Keywords: Twitter, tweets, language variation, syntactic structure, hashtag clustering

1 Introduction

These days Twitter is widely used for various purposes, from 'taking the world's pulse', i.e. using Twitter data to discover what is going on, what is occupying people's minds, etc., to gathering all sorts of information which is extremely valuable for example for marketing purposes. So far, however, most linguists appear to shun away from Twitter as a data source, possibly because the amount of data is rather daunting, but surely also because the metadata used in sociolinguistic research is usually missing. Another factor may well be that the language use found in the tweets is extremely diverse: it shows a great deal of variation (there does not appear to be such a thing as one single Twitter language), and it is unclear how it relates to the standard language as we know it and which over the past decades has been the traditional object of study in linguistics. Most studies that have investigated the language used in tweets have focused on vocabulary use and spelling variation. Beyond that, as far as we are aware, there has been no attempt to make for example a systematic inventory of the syntactic constructions that are used.

In this paper we set out on developing an approach that will eventually enable us to develop a rule-based syntactic parser for analyzing Dutch tweets. As a first step, we try to harness the wide variation by identifying the various 'tribes', i.e. communities of users

who share an interest in specific topics[1]. Our assumption is here that – as in the standard language – language variation can be attributed to two main parameters: the language user and the situational context in which he/she finds him/herself. Therefore, we cluster the tweets on the basis of the hashtags and the users that are using these hashtags. We expect that where variation in language use occurs, the variation within a cluster will be less than the variation between clusters. Next, a more detailed analysis is made of a few clusters with the aim to establish whether they lend further support to the idea that the variation observed with Twitter data can be attributed to different tribes and to uncover some of the distinctive syntactic features that characterize different tribal languages.

The remainder of this paper is structured as follows: In Section 2 we give a brief overview of related research. Then, in Section 3, we describe our approach in which we cluster the Twitter data using hashtags and users, and select four clusters for further analysis. In Section 4 we first look at what surface statistics can tell us about the saliency of the selected clusters, before proceeding (in Section 5) with the annotation of the data. In Section 6 we present the results and discuss the variation observed *between* and *within* clusters. Section 7 concludes this paper with a summary of the main findings and suggestions for further research.

2 Language variation

Research into variation in language use has a long tradition in linguistics. Numerous studies have investigated the nature of the variation and the linguistic and extra-linguistic parameters that can be associated with it. Already in early work by, for example, Gregory (1967) and Crystal & Davy (1969), it was observed that groupings of linguistics features can be correlated with recurrent situational features (cf. Gregory, 1967: 178). These situational features relate to both reasonably permanent characteristics of the user and to the recurrent characteristics of the user's use of the language. The former include the user's individuality, his temporal, geographical and social provenance, and his range of intelligibility[2]. The latter pertain to the purposive role, the medium relationship, and the addressee relationship[3]. While most previous studies before him analyzed linguistic variation in terms of a single parameter, Biber (1988) was among the first to undertake a computerized, corpus-based study using a multi-feature/multi-dimensional approach in which co-occurring linguistic features were identified empirically and interpreted in terms of their situational, functional, and processing constraints. Subsequent studies have since investigated linguistic variation both between and within genres, registers, and text types. Biber's approach has also successfully been applied to variation within texts. To date, however, most linguistic variationist studies involve more traditional data. Only more recently studies have begun to appear that include data from

[1] The term 'tribe' was taken from an article in the Guardian Datablog (15 March 2013), 'Twitter users forming tribes with own language, tweet analysis shows'.

[2] In linguistics these user-related variants are known as idiolect, temporal dialect, geographical dialect, and standard/sub-standard/non-standard language.

[3] In linguistics these underlie the distinction between the notions of field, mode and tenor of discourse.

the social media (e.g. Bamman, Eisenstein, & Schnoebelen, 2014; Puschman, 2010; Imo, 2010; Crystal, 2006).

In the field of natural language processing researchers have massively turned to the social media, as it can provide access to vast amounts of data that can be mined for all sorts of information. Often, however, the data does not reach its full potential as researchers perceive the data to be so noisy that it hinders automatic processing. They struggle with issues of encoding and syntax (in the technical sense), and mostly focus on vocabulary use, spelling, the use of emoticons, hashtags, and URLs. Variants tend to be viewed as erroneous instantiations of some standard form, usually the standard upheld in written edited published texts. Thus, Kaufmann and Kalita (2010) as regards Twitter data observe that "[T]his offers new and exciting opportunities, and there is much useful information that can be learned from meaningful analysis of the data. But the quality of the data is so poor that standard NLP tools are unable to process it."[4] According to Kaufmann and Kalita "[T]weets often contain highly irregular syntax and nonstandard use of English"[5]. In NLP, unlike in linguistics, there seem to be relatively few studies so far which consider language used on Twitter as a variety in its own right.

Studies from the field of communication studies at large have been quick to pick up on the potential of social media. However, when it comes to variation, we find that they generally look at variation from a different perspective: they are not so much interested in the linguistic features that co-occur under specific conditions, rather they focus on the communicative setting and the communicative effect of what is being communicated. The vast majority of specialized studies that target computer-mediated communication (e.g. Bieswanger, 2013; Danet, 2010; Tagg, 2009; Danet, 2001) are mostly concerned with orthography and spelling. Some studies focus on medium-specific features., like for example Shapp (2014) who investigates variation in the use of Twitter hashtags.

Community membership showing in language has been demonstrated, for example, in a recent study by Bryden, Funk, & Jansen (2014). They looked at communication on Twitter and found that "the network emerging from user communication can be structured into a hierarchy of communities, and that the frequencies of words used within those communities closely replicate this pattern" (2014:1). In their study, Bryden et al. take word usage as a proxy of variation in language. In the present study we build upon and extend the idea as we investigate the use of syntactic constructions by different communities or 'tribes'.

3 Hashtag clusters

In a first experiment we conducted, we used the TwiNL data collection, which according to the creators contains about 40% of Dutch tweets (Tjong Kim Sang & van den Bosch, 2013). We restricted ourselves to the 478,509 users who posted at least 1,000 tweets over the period January 2011 to June 2013. We then selected the 2,081 hashtags that were used by a minimum of 478 users, i.e. at least one in a thousand. We determined the pointwise mutual

[4] Cited from Kaufmann & Kalita (2010), Section II. Motivation.
[5] Cited from Kaufmann & Kalita (2010), Abstract.

information (PMI) of these users and hashtags and created a 478,509-dimensional vector for each hashtag in which each element was either the PMI value when that was positive, or 0 otherwise. We then determined a dissimilarity matrix, using 1 minus the cosine of each vector pair as the distance measure, and used this matrix to cluster the hashtags.[6]

Inspection of the output of the automatic clustering revealed that the clustering indeed appears to yield clusters of hashtags that belong to specific topics of discourse. In addition, we see clear influences of user communities. Thus we find two clusters that both have to do with 'housing'. One cluster is formed by hashtags such as #Tekoop ('For sale'), #Appartement ('Apartment'), and #Commercieel ('Commercial'), which point towards the 'selling side' of the housing market. The other cluster is formed by hashtags such as #huren ('to rent'), #makelaar ('real estate agent'), and #wonen ('to live'), which point towards the 'buying side'. Similary we see that, in the clustering of hashtags of political parties, there is a first grouping of #CU, #SGP, #GL, #Groenlinks, #D66, #PVV, #CDA, #VVD, #PvdA and #SP, and another of #sp, #groenlinks, #d66, #pvda, #vvd, #pvv and #cda before these go on to form a single larger grouping.

We selected four clusters, related to 'appreciation' (cluster A), 'employment' (cluster E), 'politics' (cluster P), and 'school' (cluster S) for further inspection. The clusters were selected on an intuitive basis: rather than making use of some threshold level, for each cluster we established the point up to where the clustered hashtags were judged to be related to a specific topical domain without including too much noise. Obviously, the intuitive judgment about the saliency of a cluster was troubled by the fact that some hashtags are more ambiguous than others. For example, #kots can be used in its literal sense ('puke'), but it can also be an expression of negative sentiment. The same is true for #blond, which literally means 'fair haired' but is quite frequently used in the sense of 'dumb' or 'not too smart', often used as a comment on the way someone acts. Yet other tags may be used in a sarcastic fashion, e.g. #bedankt ('thanks'). Still, we were fairly confident that the clusters we selected comprised sets of hashtags that were sufficiently coherent to serve as a basis for our investigation.

Cluster A which we labelled 'appreciation' comprises 74 hashtags and is by far the largest of the four clusters[7]. The binding factor for the hashtags in this cluster seems to be that they all express some kind of sentiment or opinion, such as #gaap ('yawn'), #oneerlijk ('unfair'), #onzin ('nonsense'), #spannend ('exciting').

Cluster E ('employment') comprises 21 hashtags, all of which seem to refer to things having to do with employment: #vacature ('vacancy'), #CV ('CV'), #carriere ('career'), etc. but also type of education or educational level (#MBO, i.e. medium level professional education).

The 32 hashtags that make up cluster P ('politics') refer to political parties (e.g. #CDA, #PvDA, #PVV), elections (#tk2012 and #TK2012 i.e. the 2012 election for the Tweede

[6] Using the function hclust in R (R Development Core Team, 2008).

[7] See also Appendix A where a more detailed overview is given of the composition of each of the four clusters.

Kamer 'Dutch parliament'), and political issues (e.g. #pensioen 'pension', #crisis 'crisis', and #euro 'euro').

In cluster S ('school') 18 hashtags can be found. These refer to school-related matters and range from subjects taught at school (#ak 'geography', #gs 'history', #engels 'English', #wiskunde ('maths'), to #huiswerk and #hw (both 'homework'), #wrts (a much used app for practicing vocabulary) and even social media and devices popular among pupils (#sms, #skype, #ipod).

4 Preliminary findings of syntactic variation and similarities

Before embarking on a costly annotation process, we first wanted to measure automatically whether there were any signs that the use of syntactic constructions indeed varied between the clusters. Actually measuring syntactic complexity was impossible, as there is no syntactic parser for Dutch tweets as yet. Instead, we used two word-related measures, that strictly speaking are not about syntactic structure, but that do indirectly point towards the syntactic realization of the tweets. The first measure is the tweet length in tokens. As syntactic possibilities with tweets with short lengths are limited, we can expect a higher syntactic complexity to be correlated with a greater tweet length. The second measure, the average U-score, stems from previous research (van Halteren & Oostdijk, 2015). For every word, the U-score (short for ubiquitous-score) expresses on a scale of 0 to 1 whether the word is restricted to specific topics/domains or can be used equally well with any text. As the words with high U-scores tend to be function words or other words that in the standard language are used to build larger syntactic constructions, a high average U-score over all words in a text (or here sample of tweets) indicates the presence of syntactic structures also observed in the standard language, whereas a low value may indicate highly reduced or otherwise deviating structures. In Figure 1, we plot the tweet length against the average U-score. We see clearly that the user-derived clusters, which we already validated for relation in content, also appear to show similarities in the linguistic composition of the corresponding tweets, suggesting that we can indeed proceed to the annotation of syntactic constructions.

5 Annotation

The four clusters were subjected to closer inspection. To this end we selected 20 hashtags (highlighted in Figure 1):

Appreciation (A):	#apart, #au, #blond, #gezellig, #jammer, #lekker, #leuk, #vreemd
Employment (E):	#baan, #carriere, #jobs, #productie
Politics (P):	#CDA, #crisis, #euro, #Europa
School (S):	#ak, #huiswerk, #hw, #wrts

The selection was made such that for each cluster it included hashtags spread over the cluster.

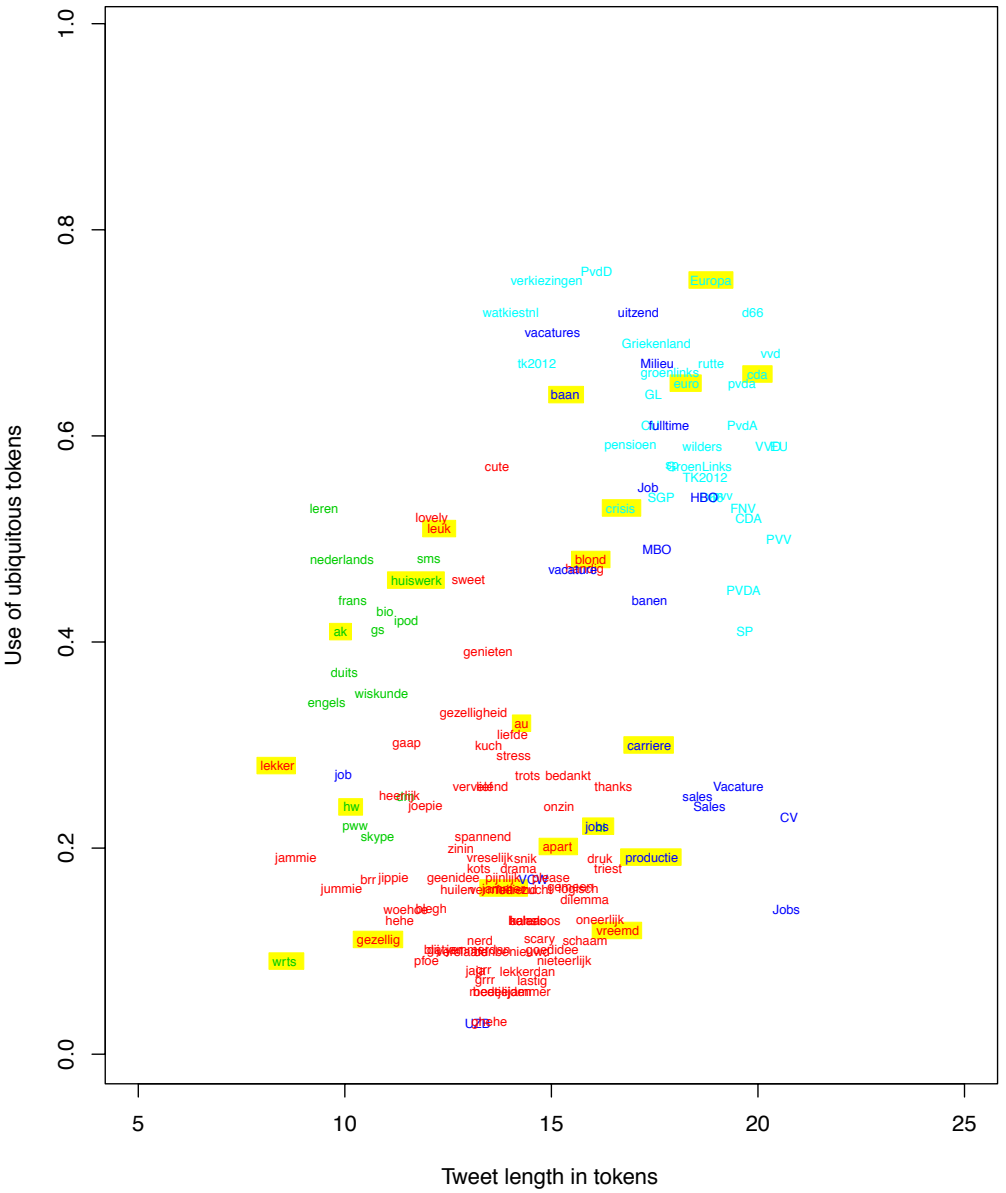


Fig. 1: Placement of hashtags from the clusters 'school' (green), 'employment' (dark blue), 'politics' (light blue), and 'appreciation' (red) with regards to tweet length and a syntax-related measurement. The hashtags highlighted with a yellow background are those for which a sample has been annotated for syntactic structure.

Table 1: **Parse units and single unit tweets per sample**. Note that the number of hashtags from which we sampled 100 tweets per hashtag is different for cluster A. There we have 8 hashtags, whereas for each of the other three clusters (E,P and S) there are only 4 hashtags.

hashtag	parse units (N)	single unit (%)	hashtag	parse units (N)	single unit (%)
#apart	140	63.0	#jammer	160	59.0
#au	151	64.0	#lekker	138	73.0
#blond	191	48.0	#leuk	174	56.0
#gezellig	151	65.0	#vreemd	167	49.0
Overall A (800)	1272	59.6			
#baan	164	69.0	#jobs	285	9.0
#carriere	265	21.0	#productie	322	2.0
Overall E (400)	1036	25.3			
#CDA	207	23.0	#euro	173	46.0
#crisis	179	44.0	#Europa	183	48.0
Overall P (400)	742	40.3			
#ak	119	86.0	#hw	155	67.0
#huiswerk	132	79.0	#wrts	137	71.0
Overall S (400)	543	75.8			

For each hashtag, a sample of 100 tweets was manually annotated. As a first step in the annotation process, tweets were (manually) split into parse units. Table 1 gives an overview of the number of parse units per sample, and for each sample, the percentage of tweets which consist of a single unit.

Next, each parse unit was annotated with information pertaining to its (surface) syntactic structure. Thus in our annotation scheme (cf. Table 2) we distinguished between various syntactic categories, viz. sentences, clauses, and phrases, as well as syntactic relations such as coordination and juxtaposition. Three additional labels were introduced. Two of these were needed to provide for parse units that were clipped or only contained some discourse item (a responsive phrase such as 'yes' or 'indeed', a greeting such as 'hi', thanks, or such like item), while the third one was used to cover any remaining realizations. While annotating the parse units for syntactic information, user names, emoticons, hashtags, and URLs were ignored, unless they were an integral part of the syntactic structure, as for example when a hashtag featured as a word in a sentence or phrase.

We also annotated for the occurrence of dependent clauses. Apart from recording their frequency of occurrence, dependent clauses were labelled for their syntactic function: utterance (UTT), subject (SU), notional subject (NOSU), direct object (OD), adverbial (A), preposi-

[8] The annotation scheme is described in some more detail in Appendix B. There also examples are given of how the labels apply.

Table 2: **Annotation scheme**[8]

label	**denotation**
coordS	coordinated sentences
juxtS	juxtaposed sentences
S(decl)	declarative sentence
S(decl,-su)	declarative sentence with subject missing
S(decl,-su,-aux)	declarative sentence with subject and auxiliary missing
S(decl,-su,-cop)	declarative sentence with subject and copular verb missing
S(decl,-su,-v)	declarative sentence with subject and verb missing
S(decl,-v)	declarative sentence with verb missing
S(decl,-od)	declarative sentence with direct object missing
S(int)	interrogative sentence
S(int,-su,-aux)	interrogative sentence with subject and auxiliary missing
S(int,-cop)	interrogative sentence with copular verb missing
S(decl/int)	declarative or interrogative sentence (undecided)
S(excl)	exclamatory sentence
S(imp)	imperative sentence
CL	clause
NP	noun phrase
AJP	adjective phrase
AVP	adverb phrase
PP	prepositional phrase
VP	verb phrase
DISC	discourse item (word or phrase)
CLIPPED	clipped uni
REST	other (none of the above)

tional complement (PC), noun phrase postmodifier (NPPO), adjective phrase postmodifier (AJPO), or adverb phrase postmodifier (AVPO)[9].

6 Results and discussion

In this section we discuss the nature of the variation that we see in the use of syntactic constructions in our data. We first show that the four clusters are again recognizable by considering all constructions together by way of a principal component analysis (Section 6.1). We then look at some individual constructions, discussing the variation observed both between clusters (Section 6.2) and within clusters (Section 6.3).

But before moving on to the investigation of the clusters, we would like to point out that the observed constructions, except for CLIPPED (7.2%) and some of REST (3.8%), are all part of the inventory of constructions found in the standard language. It is merely their use that is different. However, the constructions are often strung together without using

[9] For examples, see Appendix B.

the punctuation that we would expect in written language. This means that, if we are to build a syntactic parser for Dutch tweets, we will have to either find a way to identify the individual parse units automatically, or have our grammar describe sequences of parse units. For standard written texts, the latter option can be expected to lead to major computational problems, but, given the length restriction on tweets, we think it could be applied in this case.

6.1 Principal component analysis

We performed a principal component analysis (PCA; Pearson, 1901; Hotelling, 1930)[10] on the thirteen characteristics for each hashtag shown in the biplot in Figure 2. These characteristics are mostly the frequencies for the constructions listed in Table 2. However, we merged some rare classes. The labels S_decl_ellip and S_int_ellip cover all types of declarative and interrogative sentences with ellipted elements respectively, and the label phrase covers all phrase types: NP, AJP, AVP, PP and VP. Similarly, we did not distinguish between various types of dependent clauses, but included all in a single characterictic DEPCL. Finally, we added the characteristic NUMUNITS, which represents the number of parse units per 100 tweets, as listed in Table 1.

In the biplot, our selected clusters are again concentrated to some degree in specific areas of the plot. The first component differentiates mostly between the employment cluster (dark blue) and the other three, with the employment direction linked to larger numbers of parse units, realized mostly by phrases, and higher frequencies of clipped tweets, whereas the other three show lower numbers of parse units and more clausal constructions. On the second component, we see a high placement of the politics cluster, showing remarkably high frequencies of interrogative and imperative sentences, and a low placement of the school cluster, with ellipsis as its most prominent characteristic. The appreciation cluster is rather spread out in these two dimensions, but clearly on the left side with politics and school.[11]

6.2 Language use and variation between the four clusters

When we look at how the four clusters compare to each other (cf. Table 3 and Figure 3), we find that

– In three of the four clusters (A, P and S) the declarative sentence is used quite frequently. As shown in Table 3, declarative sentences account for 65.8% (cluster A), 49.9% (cluster

[10] A principal component analysis rotates the data space in such a way that the variance in the data is shown as much as possible in the earlier dimensions, i.e. the first dimension covers the highest possible proportion of variance, the second the highest possible proportion within the remaining variance, etc. After ordering the dimensions in this way, we can choose to represent the data by a lower number of dimensions while keeping as much information as possible (dimensionality reduction). A plot of the first two dimensions (principal components) is thus assumed to show the most interesting aspects of the data. In such a plot, we can also show where the originally measured characteristics would be placed; this leads to a so-called biplot (e.g. Figure 2).

[11] The appreciation cluster is separated from politics and school only in the third and fourth principal component.

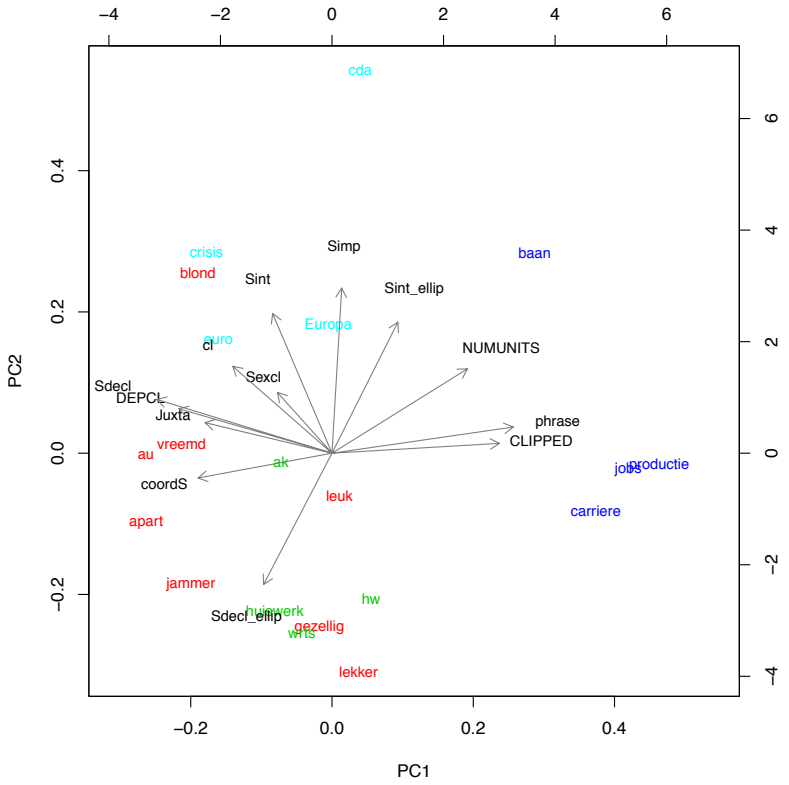


Fig. 2: Biplot showing the placement of the hashtags and the individual features in relation to the first two principal components. The hashtags are again marked with the cluster colours: 'school' (green), 'employment' (dark blue), 'politics' (light blue), and 'appreciation' (red).

P) and 70.5% (cluster S) of the units. Cluster E with a mere 7.4% of the units realized by a declarative sentence stands out.

- What is striking is that in cluster S the majority of the declarative sentences is elliptic (50.8% vs 19.7%), whereas in cluster P most declarative sentences are full sentences. In cluster A, full and elliptic sentences occur equally frequently (32.9%).
- Interrogative sentences are most frequent in cluster P (10.4%) and S (6.1%).
- Imperative sentences on the whole are rare, except for cluster P where they account for 6.3% of the units.

Table 3: **Linguistic syntactic structure: Four clusters**. Figures denote the percentage of parse units that were assigned these labels.[12]

Cluster	S(decl)	S(decl,ellip)	S(int)	S(imp)	phrase
Appreciation	32.9	32.9	3.5	1.3	10.3
Employment	4.5	2.9	1.3	1.4	67.4
Politics	34.5	15.4	10.4	6.3	16.6
School	19.7	50.8	6.1	0.6	9.9

– In cluster E the largest proportion of units by far (67.4%) is realized by a phrase; in none of the other clusters are phrases that frequent.

– To the extent that dependent clauses are used, they mainly occur in clusters A and P, more specifically as direct object, adverbial and noun phrase postmodifier.

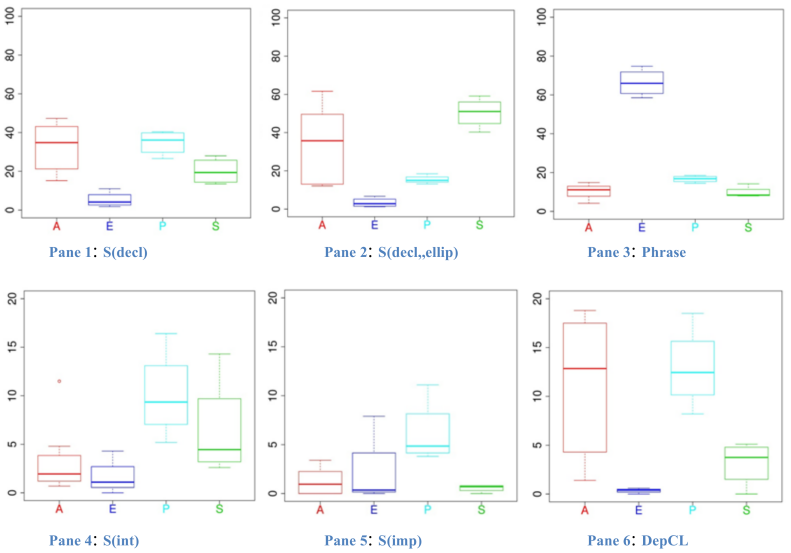


Fig. 3: Boxplots showing the frequencies of specific constructions (Panes 1 to 6) in each cluster: 'school' (green), 'employment' (dark blue), 'politics' (light blue), and 'appreciation' (red). Note that the vertical axis has a different scale in Panes 1 to 3 (0-100%) from that in Panes 4 to 6 (0-20%).

[12] The table only includes the main distinctive features.

6.3 Language use and variation within each of the clusters

As we saw in Section 6.2, the four clusters are quite different from each other. There is, however, also considerable variation within the clusters (cf. Figure 3). Below we discuss each of the clusters in some more detail.

Appreciation. For all of the 8 hashtags we inspected in the appreciation cluster, the most common realization of a unit is the declarative sentence (cf. Table 4). There is, however, quite some variation among these hashtags. In the case of #gezellig and #lekker, elliptic sentences appear to be favored over full declarative sentences (56.7 and 61.6 percent of elliptic sentences respectively), whereas with #blond, #apart and #vreemd the opposite is true: here we find only 12.1, 12.9 and 13.9 percent of elliptic sentences respectively. In this cluster, only for #vreemd do we find quite a large proportion of units that appear to be clipped (7.8%). These are typically found in tweets where news facts are commented upon as in examples [1] and [2].

[1] #vreemd Britse premier Cameron vergeet dochter in pub: De Britse premier David Cameron heeft zijn 8-jarige docht... `http://t.co/Ekvx7wME`
EN: British prime minister Cameron forgets daughter in pub. The British prime minister David Cameron has [...] his 8-year-old daught...
(hashtag: odd)

[2] #vreemd Vrouw stript voor geld langs kant van de weg: AMSTERDAM - In de Amerikaanse staat Pennsylvania is de 35-... `http://t.co/71teCBIa`
EN: Woman strips for money by the side of the road: AMSTERDAM — In the American state Pennsylvania the 35-...
(hashtag: odd)

Most tweets in the appreciation cluster are about expressing some sentiment or opinion on some stated fact or observation (hence the frequent use of declaratives). The overall proportion of interrogatives then may seem somewhat surprising. However, as the figures in Table 4 show, interrogatives are especially frequent with #blond. In this cluster typically questions are asked about how things work or to have something clarified, using the 'being blond' as an explanation for (some)one's ignorance (exs. [3]-[4]).

[3] waarom doet me #htc geen internet zonder wi-fi? en hoe zet ik dat goed #blond #durftevragen
EN: why doesn't my #htc have internet without wi-fi? and how can I correct that
(hashtags: not very smart, dare to ask)

[4] Waarom hebben ze het over een 'definitieve prognose'? #tk2012 #blond?
EN: Why are they talking about a 'definitive prognosis'?
(hashtags: parlement2012, not very smart)

In other cases where #blond is found, it is used to comment on someone (quite often the user him- or herself) doing something stupid (exs. [5]-[6]).

Table 4: **Linguistic syntactic structure: Appreciation cluster**. Figures denote the percentage of parse units that were assigned these labels. In each column the highest and lowest percentages are indicated using bold face type (underlined and italics respectively).

Hashtag	**S(decl)**	**S(decl,ellip)**	**S(int)**	**S(imp)**	**phrase**
#apart	46.4	12.9	2.9	0.0	10.7
#au	35.8	37.1	2.6	0.0	5.3
#blond	39.8	12.1	11.5	3.1	12.6
#gezellig	16.6	56.7	1.3	1.3	13.3
#jammer	33.8	34.4	1.3	0.0	10.3
#lekker	15.2	61.6	0.7	1.4	11.6
#leuk	25.9	42.5	1.1	3.4	14.9
#vreemd	47.3	13.2	4.8	0.6	4.2

[5] Ik heb me hoesje verkeert er omheen gedaan #blond
EN: I put on the cover the wrong way round
(hashtag: not very smart)

[6] Bedenk me net dat ik vergeten ben uit te checken met de ov-chipkaart. #handig #duurgeintje #blond #gelukkigstaaternietzoveelmeerop
EN: I just realized that I forgot to check out with the Oyster card.
(hashtags: smart, costly affair, not very smart, fortunately there was not much money on it any more)

Noteworthy for #blond is also the large proportion (6.8%) of juxtaposed sentences. A typical examples are given in [7] and [8]. With all other hashtags (in this cluster but also in the other clusters) juxtaposed sentences were quite rare.

[7] ik dnek wanneer begint mn liedje nou, staat mn geluid uit ! #blond ?
EN: I'm wondering when will my song start, my sound is switched off !
(hashtag: not very smart)

[8] ik ben bekant bij de sportschool. staat edwin achter me. ik was mn id vergeten. #blond?
EN: I'm almost at the sports school. Edwin turns up behind me. I'd forgotten my ID.
(hashtag: not very smart)

Finally, whereas coordinated sentences are rare in all clusters and hashtags, #apart, #jammer, and #vreemd have exceptionally large proportions of coordinated sentences: 12.1, 10.6, and 12 percent of the units respectively are coordinated sentences.

Employment. The tweets in the employment cluster are mostly from professional/commercial users. In this cluster #baan is typically the odd one out: #baan has by far the largest proportion of tweets (69/100) that are composed of a single unit.[13] Moreover, tweets that occur with #baan stand out in this cluster as they, compared to the other hashtags, contain sentences relatively more frequently (cf. Table 5). Typical examples from #baan are [9] to [11].

[13] Cf. #carriere 21/100, #jobs 9/100, and #productie 2/100 (Table 1).

[9] PostNL zoekt Postbezorgers (M/V) in eigen wijk in Woerden of omgeving: http://t.co/mndPG6bU #vacature #postbode #baan #bijbaan #werk
EN: PostNL is looking for mailmen (M/F) in own district in Woerden or surroundings
(hashtags: vacancy, mailman, job, work)

[10] Sales Representative Retail http://t.co/E9BXmnWZ #werk #werken #banen #baan #vacature #vacatures #MBO #HBO #WO #job #jobs #werkzoeken #in
EN: Sales representative retail
(hashtags: work (noun), work (verb), jobs, job, vacancy, vacancies, professional education, higher professional education, university education, job, jobs, looking for work, in)

[11] Milieuopleiding afgerond? Vind je eerste #baan bij EcoFlexx! http://t.co/PCneSzum #Milieu #Bodem #Water #Civiel #RO #Job #Vacature
EN: Environmental training completed? Find your first job with EcoFlexx!
(hashtags : environment, soil, water, civil, spatial planning, job, vacancy)

Especially with #carriere, #jobs and #productie we find that tweets are composed of multiple units, and the units appear to be ordered according to a fixed pattern (although there is some variation among the various users – MonsterBanen, Megajobs, JobshopNL,...; exs. [12]-[13]). This suggests the use of some kind of template. Our impression is that it may be well possible that the tweets were generated automatically using a database listing the job title, location, company, and job description.

[12] JobshopNL #Jobs ? Toetdeskundige Wiskunde rekenen: De functie Als toetsdeskundige begeleid... http://t.co/RboGj424tT #vacature #carriere
EN: JobsNL ? exam expert Maths arithmetic: The function As exam expert [you] guide ...
(hashtags: Jobs, vacancy, carreer)

[13] #vacatures #carriere Engineer Sprinklertechniek: OV-Harderwijk, Als ervaren Engineer Sprinklertechniek en... http://t.co/SDvI4DC9 #banen
EN: Engineer Sprinkler technology: OV-Harderwijk, As experienced Engineer Sprinkler technology and ...
(hashtags: #vacancies, career, jobs)

With #carriere we find some tweets that originate from private users. Typically, these are not about employment, but mostly about the computer game fifa (exs. [14]-[15]).

[14] Wejowejo, ik gooi Bony erin, minuut later staat 't 2-0. #carriere #vitesse
EN: I send in Bony, a minute later it's 2-0.
(hashtags: career, vitesse)

[15] Erg zuur hoor... marcelo en Hulk zijn geblesseerd geraakt #psv #carriere #fifa13
EN: quite a shame ... marcelo and Hulk got injured
(hashtags: psv, carreer, fifa13)

As we noted above, the overall proportion of clipped units in this cluster is substantial: 20.1%. The proportions for #carriere, #jobs and #productie can be said to be comparable (23.8%, 18.2% and 25.2% resp.), while for #baan it is only 7.8%.

Table 5: **Linguistic syntactic structure: Employment cluster**. Figures denote the percentage of parse units that were assigned these labels. In each column the highest and lowest percentages are indicated using bold face type (underlined and italics respectively).

Hashtag	**S(decl)**	**S(decl,ellip)**	**S(int)**	**S(imp)**	**phrase**
#baan	11.0	6.7	4.3	7.9	58.5
#carriere	4.9	3.8	1.1	0.0	63.0
#jobs	1.8	1.8	1.1	0.4	74.7
#productie	3.4	1.2	0.0	0.3	68.9

Politics. In the politics cluster we (perhaps naively) expected to find tweets related to political parties, elections, political issues, etc. For #CDA this is mostly the case. Many tweets appear to originate from professional users (e.g. PartijNieuws, CDAKrant; exs. [16]-[17]).

[16] #CDA D66, VVD en CDA voor extra geld voor Strafhof – Eindhovens Dagblad `http://t.co/jQunEwY0`
EN: D66, VVD and CDA pro extra funds for Judicial Court — Eindhovens Dagblad[14]
(hashtag: CDA)

[17] #CDA Buma (CDA) sluit samenwerking met PVV uit: DEN HAAG — Het CDA zal na de verkiezingen van 12 september niet ... `http://t.co/9ss9gAB3`
EN: Buma excludes collaboration with PVV: The Hague -– After the elections of September 12th the CDA will not ... (hashtag: CDA)

However, there are also tweets that are only marginally related to politics. This is the case, for example, for the user VakmensenNL, that uses #CDA merely to be found by search engines. A tweet from VakmensenNL typically has two units, the first of which is an interrogative sentence, the second an imperative, as in examples [18] and [19].[15]

[18] Stem u op #CDA? Kijk dan bij de #Voedingsvoorlichter - #verkiezingen - `http://t.co/2Y099s6ZZl`
EN: Are you voting CDA? Look at Nutrition advisor
(hashtags: nutrition advisor, elections)

[19] Stem u op #CDA? Kijk dan bij de #Apotheker – #stemmen – `http://t.co/c7EgtTsa`
EN: Are you voting CDA? Look at Chemist's
(hashtag: chemist's, vote)

The other hashtags in this cluster are even more noisy:

– 30/100 tweets with #Europe are not about politics at all. They are about a variety of topics: football, traveling, sales, etc.

[14] D66, VVD, CDA, and PVV are Dutch political parties. Buma is a Dutch politician. Eindhovens Dagblad is a daily newspaper.

[15] This may well explain the even for the politics cluster large proportions of interrogative and imperative sentences.

Table 6: **Linguistic syntactic structure: Politics cluster**. Figures denote the percentage of parse units that were assigned these labels. In each column the highest and lowest percentages are indicated using bold face type (underlined and italics respectively).

Hashtag	**S(decl)**	**S(decl,ellip)**	**S(int)**	**S(imp)**	**phrase**
#CDA	26.6	15.0	16.4	11.1	14.5
#crisis	33.0	15.1	8.9	4.5	16.2
#Europa	40.4	13.2	9.8	3.8	17.4
#euro	39.3	18.5	5.2	5.2	18.5

– about 1 in 5 tweets with #euro are not about the euro as a political issue, but are rather about promotional actions and sales offers (ex. [20]).

[20] #deal #www Doe mee met deze #actie en maak #kans op 150 #euro #tank tegoed `http://t.co/nfdiSFZfb7` #prijsvraag.
EN: Participate in this promotion and stand a chance to win 150 euro worth of petrol
(hashtags: deal, www, promotion, chance, euro, tank, contest)

– although the majority of tweets with #crisis is about the economic crisis (exs. [21]-[22]), there are also a substantial number of tweets about some 'crisis' in someone's personal life (exs. [23]-[24]):

[21] 'Meer duurzame woningen bouwen om crisis te lijf te gaan' `http://t.co/9I5oMm9l` #duurzaamheid #crisis #woningbouw
EN: 'Build more sustainable houses to fight crisis'
(hashtags: sustainability, crisis, building houses)

[22] #crisis #nieuws: EU: Nederlands tekort 3,6 procent, economie krimpt met 0,6 procent `http://t.co/S0hjDocuCw`
EN: EU: Dutch deficit 3.6 per cent, economy shrinks with 0.6 per cent
(hashtags: crisis, news)

[23] @gijsvofoto Kat naar utrecht voor operatie aan pootje? Dat meen je niet! #crisis
EN: Cat to Utrecht for operation on leg? You're kidding!
(hashtag: crisis)

[24] Weeh ik kan me bb niet meer vinden ? #crisis
EN: Sob I can't find my blackberry anymore?
(hashtag: crisis)

School. The tweets in this cluster reflect the use of the medium as a means for high school kids to interact with their peers. The communication is typically bi-directional as they are sharing information about what they are doing and voicing opinions, but also from time to time asking questions (e.g. about what homework assignments they have got) and inviting responses from their mates. There are many short exchanges, but from time to time there are also longer tweets which then typically are very dense, packed with information, and highly elliptical (cf. exs. [25]-[28]).

Table 7: **Linguistic syntactic structure: Politics cluster**. Figures denote the percentage of parse units that were assigned these labels. In each column the highest and lowest percentages are indicated using bold face type (underlined and italics respectively).

Hashtag	**S(decl)**	**S(decl,ellip)**	**S(int)**	**S(imp)**	**phrase**
#ak	23.5	40.3	14.3	0.8	8.4
#huiswerk	28.0	49.2	3.8	0.8	8.3
#hw	13.5	52.9	2.6	0.6	14.2
#wrts	15.3	59.1	5.1	0.0	8.0

[25] Wiskunde maken, ofso ? #huiswerk
EN: Do maths or something?
(hashtag: homework)

[26] Pff kut techniek #huiswerk
EN: Pff damn technology
(hashtag: homework)

[27] @sannelepoutre ooh oke superleuk! ja dat is minder. is het veel dan? ik moet ook leren in de vakantie anders ken ik 't niet ... #ak
EN: @sannelepoutre oh okay supercool! yes that sucks. is it much then? I also have to study during the holidays otherwise I won't know it ...
(hashtag: geography)

[28] En weer thuis. Lekker gesport vandaag. Woensdag is de dag van vele uurtjes. Zometeen verslag vertalen. Duits & Nederlands #huiswerk
EN: And home again. Great work out today. Wednesday is the day of many hours. In a minute translate report. German & Dutch
(hashtag: homework)

When we look at the variation within the cluster (cf. Table 7), we see that what was already predicted by the U-scores (cf. Section 4) is confirmed: with #huiswerk and #ak (both relatively high U-scores) more traditional structures are found, whereas with #hw and #wrts (relatively low U-scores) there is indeed frequent use of reduced or otherwise deviating structures.

In this cluster the large proportion of interrogative sentences found with #ak and of phrases with #hw beg an explanation. The hashtag #ak is the only one in our selection that represents an actual school subject. The interrogative sentences are mostly of two types, connected to this fact. First there are questions about which material was discussed and what homework needs to be done (exs. [29]-[30]).

[29] moet je de schaal van richter kennen ? #ak #tk2
EN: Do we have to know Richter's scale ?
(hashtags: geography, tk2)

[30] Tot hoe ver moet je lezen #ak ?

EN: Up to what point do you have to read ?
(hashtag: geography)

Secondly, there are questions about the material itself, such as the meaning of specific terms (ex. [31]).

[31] Wat zijn lithosferische platen? #ak
EN: What are lithospheric plates?
(hashtag: geography)

In the case of #hw what we observe is that this hashtag is quite noisy: apart from the school-related texts we find many news items from a local weekly newspaper (viz. Harener Weekblad). Many of the phrases (14.2%) are in fact the name of a location, such as Groningen in the examples below (exs. [32]-[33]).

[32] #hw Politie onderzoekt twee straatroven – GRONINGEN – De politie stelt een onderzoek in naar twee straatroven ...`http://ow.ly/1ea52v`
EN: Police investigates two street robberies – GRONINGEN – The police are investigating two street robberies...

[33] #hw Man zwaar gewond bij schietincident – GRONINGEN – Een 24-jarige man uit Paterswolde is dinsdagavond op de ...`http://ow.ly/1e4OKR`
EN: Man severely injured during shooting incident – GRONINGEN – Tuesday night a 24-year-old man from Paterswolde was on the

7 Conclusion

In this paper we set out to investigate the language used on Twitter, with a focus on syntax. We wanted to see to what extent it varies and how it relates to the standard language. The main findings of the exploratory research reported here can be summarized as follows:

- After clustering a collection of tweets on the basis of users and hashtags, and closer inspection of four of the resulting clusters, we found that each of the four clusters indeed bears characteristics that set it apart from the other clusters. As was already hinted at by the U-scores, the clusters exhibit variation in language use which is also manifest in the linguistic structures that are used. A tentative explanation for the variation *between* the clusters is that with clusters E and P the language use is typically informative, whereas in the A and S clusters it is highly interactional. This, however, needs further investigation.
- There is variation *within* clusters. The variation is often related to the specific topics discussed with the various hashtags and/or to specific users or user groups. Also, part of the variation could be explained through the noisiness of the hashtags, which played a role for some hashtags more than for other ones.
- Language use in terms of syntactic constructions is quite regular: it essentially makes use of the same inventory of constructions that we find in standard language; to the extent that Twitter tribal languages differ from other varieties, the difference is in the frequency and distribution of various constructions.

In this paper the focus has been mainly on the composition of tweets in terms of parse units and their syntactic structure. There is much more to be said about the use of dependent clauses and the structure of phrases. For example, dependent clauses appear to be heavily underused when compared to more conventional text types. To the extent that they do occur, it is mostly as adverbial or direct object. Also, while annotating the parse units, we observed that frequently in noun phrases determiners are omitted and postmodification is much simplified or lacking altogether. These are all indications that the language is being adapted to the situational context, not least the Twitter medium.

As for the feasibility of a (rule-based) syntactic parser for Dutch tweets, based on what we have seen in this exploratory study we believe that the development of such a parser is very well possible. The grammar rules by themselves can be derived directly from existing grammars for standard written language, as long as these include proper treatment of ellipsis. If a probabilistic component is used for disambiguation, probabilities would best be recalculated per tribe, but this is still doable. The main problem for the parsing lies in the steps that precede the syntactic analysis, especially for the more informally and intimately tweeting tribes, namely dealing with the deviations from standard use of punctuation and extensive spelling variation.

Bibliography

[1] Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender and variation in social media. *Journal of Sociolinguistics*, 18.2, 135-160. http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/abstract

[2] Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

[3] Bieswanger, M. (2013). Micro-linguistic structural features of computer-mediated communication. In S. Herring, D. Stein, & T. Virttanen (Eds.), *Pragmatics of Computer-Mediated Language* (pp. 463-486). Berlin/Boston: De Gruyter.

[4] Bryden, J., Funk, S, & Jansen, V. (2013). Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, 2:3. doi: 10.1140/epjds15

[5] Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

[6] Crystal, D., & Davy, D. (1969). *Investigating English Style*. London: Longmans.

[7] Danet, B. (2001). *Cyberpl@y: Communicating Online*. Oxford: Berg.

[8] Danet, B. (2010). Computer-mediated language. In J. Maybin & J. Swann (Eds.). *The Routledge Companion to English Language Studies* (pp. 146-156). Abingdon, Oxon: Routledge.

[9] Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics*, 3, 177-198.

[10] Guardian Datablog (15 March 2013). Twitter users forming tribes with own language, tweets analysis shows. Retrieved from http://www.theguardian.com/news/datablog/2013/mar/15/twitter-users-tribes-language-analysis-tweets

[11] Halteren, H. van, & Oostdijk, N. (2015).Word distributions in Dutch tweets: A quantitative appraisal of the distinction between function and content words. *Tijdschrift voor Nederlandse Taal- en Letterkunde, TNTL*, pp. 189-226

[12] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, and 498–520.

[13] Imo, W. (2010). How syntactic structures and genres interact. In H. Dorgeloh & A. Wanner (Eds.). *Syntactic Variation and Genre. Topics in English Linguistics* (pp. 141-166). Berlin/New York: Mouton De Gruyter.

[14] Kaufmann, M., & Kalita, J. (2010). Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India. http://www.cs.uccs.edu/~jkalita/work/reu/REUFinalPapers2010/Kaufmann.pdf

[15] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2 (11), 559–572.

[16] Puschmann, C. (2010). Thank you for thinking we could: Use and function of interpersonal pronouns in corporate web logs. In H. Dorgeloh & A. Wanner (Eds.). *Syntactic Variation and Genre. Topics in English Linguistics* (pp. 167-191). Berlin/New: Mouton De Gruyter. York.

[17] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org

[18] Shapp, A. (2014). *Variation in the Use of Twitter Hashtags*. Qualifying paper in sociolinguistics. New York: New York University.

[19] Tagg, C. (2009). *A Corpus Linguistics Study of SMS Text Messaging*. Birmingham: University of Birmingham.

[20] Tjong Kim Sang, E. and Bosch, A. van den (2013). Dealing with Big Data: The case of Twitter. *CLIN Journal* Vol. 3, 121-134.

A Clusters

Below an overview is given of the composition of each of the four clusters that feature in the present study. The selective hashtags that we subjected to further analysis are marked in bold face type.

In figures A.1 to A.4 excerpts of the dendrogram for the cluster structure are shown. Each figure visualizes the composition of one of the four clusters, including only the hashtags selected for annotation. The top node represents the point where the cluster joins another node in the dendrogram. The uninterrupted lines indicate a direct connection; the dashed lines indicate an indirect connection, while the hashtag dominated by a dashed line is one of a group of hashtags that cluster at that point.

Table A.1: **Appreciation** (cluster of 74 hashtags; Figure A.1).

#apart	#gezellig	#kuch	#spannend
#au	#gezelligheid	#lastig	#stress
#balen	#ghehe	#lekker	#sweet
#bedankt	#goedidee	#lekkerdan	#thanks
#beetjejammer	#grr	#leuk	#triest
#benbenieuwd	#grrr	#lief	#trots
#blegh	#handig	#liefde	#vermoeiend
#blij	#heerlijk	#logisch	#verslaafd
#blond	#hehe	#lovely	#vervelend
#brr	#helaas	#medelijden	#vreemd
#cute	#huilen	#nerd	#vreselijk
#dilemma	#jaja	#nieteerlijk	#woehoe
#drama	#jammer	#oneerlijk	#zinin
#druk	#jammerdan	#onzin	#zucht
#feit	#jammie	#pfoe	
#gaap	#jippie	#pijnlijk	
#gatver	#joepie	#please	
#geenidee	#jummie	#scary	
#gemeen	#kansloos	#schaam	
#genieten	#kots	#snik	

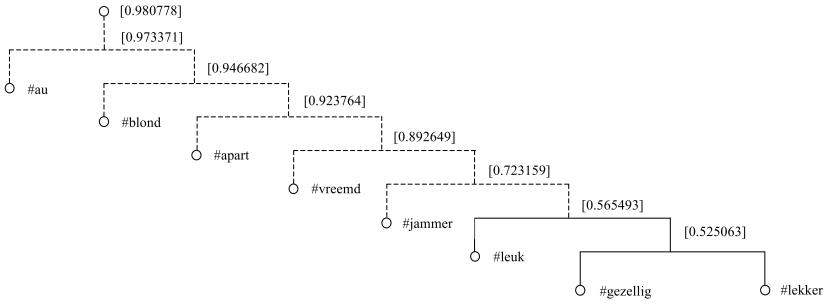


Fig. A.1: Excerpt of the dendrogram showing the selected hashtags from the 'appreciation' cluster.

Table A.2: **Employment** (cluster of 21 hashtags; Figure A.2).

#baan	#job	#productie	#vacature
#banen	#Job	#Sales	#vacatures
#carriere	#Jobs	#sales	#VCW
#CV	#jobs	#uitzend	
#fulltime	#MBO	#UZB	
#HBO	#Milieu	#Vacature	

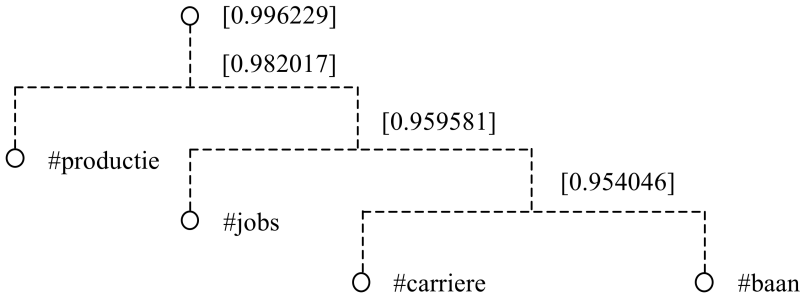


Fig. A.2: Excerpt of the dendrogram showing the selected hashtags from the 'employment' cluster.

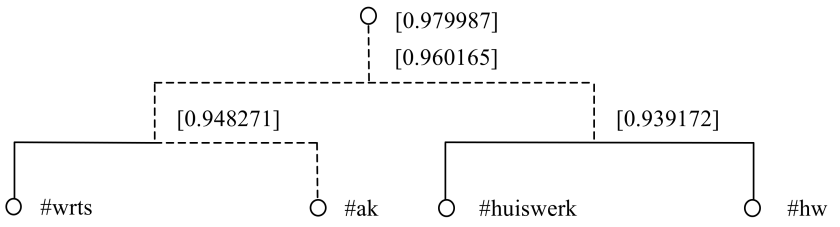


Fig. A.4: Excerpt of the dendrogram showing the selected hashtags from the 'school' cluster.

Table A.3: **Politics** (cluster of 32 hashtags; Figure A.3).

#CDA	#Europa	#PVDA	#SP
#cda	#FNV	#PvdA	#tk2012
#crisis	#GL	#PvdD	#TK2012
#CU	#Griekenland	#PVV	#verkiezingen
#d66	#groenlinks	#pvv	#VVD
#D66	#Groenlinks	#rutte	#vvd
#EU	#pensioen	#SGP	#watkiestnl
#euro	#pvda	#sp	#wilders

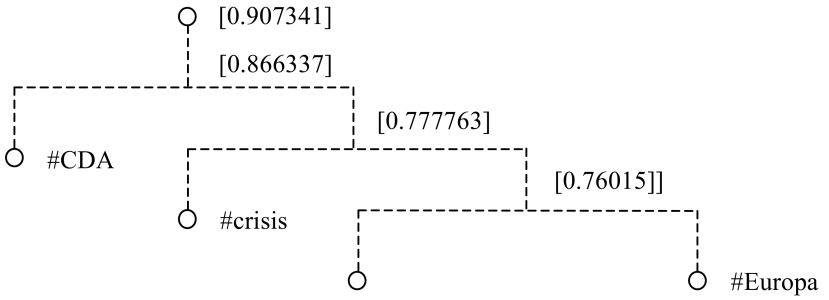


Fig. A.3: Excerpt of the dendrogram showing the selected hashtags from the 'politics' cluster.

Table A.4: **School** (cluster of 18 hashtags; Figure A.4).

#ak	#frans	#leren	#sms
#bio	#gs	#nederlands	#wiskunde
#dm	#huiswerk	#nl	#wrts
#duits	#hw	#pww	
#engels	#ipod	#skype	

B Annotation scheme

Below we repeat the annotation scheme that we presented in Table 2, this time with examples. Most examples were encountered as original tweets; occasionally an example did not constitute a tweet by itself but was identified as one of multiple parse units making up a tweet.

label	denotation
coordS	coordinated sentences
	Wow k loop nr buiten en daar staat een fels malibu :$:p #apart
	EN: Wow I walk outside and there is a bottle of Malibu :$:p #strange
juxtS	juxtaposed sentences
	ik dnek wanneer begint mn liedje nou, staat mn geluid uit ! #blond ?
	EN: I think when will my song start, my sound is muted ! #notverysmart ?
(decl)	declarative sentence
	#SPnl #Europa - @DdJong: Europese Commissie wil wurgcontracten opleggen aan lidstaten: Phttp://t.co/w6oahBQc60 via @spnl
	EN: #SPnl #Europe - @DdJong: European Commission wants to impose killer contracts on member states `http://t.co/w6oahBQc60` via @spnl
S(decl,-su)	declarative sentence with subject missing
	ben echt bang voor morgen #ak #laatste #test
	EN: am really afraid for tomorrow #geography #last #test
S(decl,-su,-aux)	declarative sentence with subject and auxiliary missing
	filmpje over aardbevingen kijken #ak
	EN: watch film about earthquakes #geography
S(decl,-su,-cop)	declarative sentence with subject and copular verb missing
	ng bij me neefje thuis #gezellig
	EN: still at my cousin's #enjoyable
S(decl,-su,-v)	declarative sentence with subject and verb missing
	niks te doen deze zaterdagavond #jammer
	EN: nothing to do this Saturday evening #pity
S(decl,-v)	declarative sentence with verb missing
	Een sp- stemmer en een pvv-er samen op 1 bank, lol
	EN: An SP voter and PVV supporter together on one couch, lol
S(decl,-od)	declarative sentence with direct object missing
	@RedCapMusic Ik weet niet man
	EN: @RedCapMusic I don't know man
S(int)	interrogative sentence
	Wat is er nodig om #positiefjeugdbeleid méér als basis te krijgen voor lokaal beleid en decentralisatie #peerlearing #Europa
	EN:What is needed to get #positiveyouthpolicy as basis for local policy and decentralisation #peerlearning #Europe
S(int,-su,-aux)	interrogative sentence with subject and auxiliary missing
	Misschien gewoon om de beurt in deeltijd?
	EN: Maybe just one at a time parttime?
S(int,-cop)	interrogative sentence with copular verb missing
	@Miekje25 iemand met reservesleutel in de buurt? #blond
	EN: @Miekje25 anyone around with a spare key? #notverysmart

label	denotation
S(decl/int)	declarative or interrogative sentence (undecided) *ik heb mijn tv aangezet in mijn slaap?!?!?* EN: I have turned on my tv in my sleep?!?!?!
S(excl)	exclamatory sentence *Wat doet/acteert die Peetoom vreselijk overdreven!* EN: How exorbitant does that Peetoom behave/act!
S(imp)	imperative sentence *Lees het gehele artikel trouwens hier:* `http://t.co/2gSg8yQx` *#Krugman #euro* EN: Read the full article here by the way: `http://t.co/2gSg8yQx` #Krugman #euro
CL	clause *@RonKas1 @rjvanhouten @adresonbekend als het programma ondertussen dan maar niet is wegbezuinigd #crisis* EN: @RonKas1 @rjvanhouten @adresonbekend If only the programme meanwhile has not been stricken off because of financial cutbacks #crisis
NP	noun phrase *Nieuw Blog Bericht* EN: New blog post
AJP	adjective phrase *suupergooooed!* EN: fabulous!
AVP	adverb phrase *thuis #hw ;s* EN: home #homework ;s
PP	prepositional phrase *Net als die andere: #PvdA* EN: Just like the other: #PvdA
VP	verb phrase *leren #ak #twexit* EN: study #geography #twexit
DISC	discourse item (word or phrase) *Huh?* EN: Eh?
CLIPPED	clipped unit *Ombouwen en afstellen van afv...* `http://t.co/oVHcCU6Z` *#techniek #productie* EN: Converting and tuning of (incomplete)... `http://t.co/oVHcCU6Z` #technology #production
REST	other (none of the above) *hoppa!* EN: hoppa!

#TweetCommeUneFille: Twitter as giveaway for stereotyping speech acts

Camille Lagarde-Belleville and Michel Otell

Department of Language Sciences
Praxiling UMR 5267-CNRS
Paul Valéry University
Montpellier, France

Keywords: Discourse Analysis, Cognitive Semantics, Gender Stereotypes, Categorisation, Digital Utterance.

1 Introduction

This work aims to tackle a corpus stemming from Twitter, a social media that is representative of emerging trends in speech. The corpus here to be studied was selected and gathered through a keyword-like tagging device: the hashtag. In 2013, Marie-Anne Paveau referred to hashtags as *technomorphemes* that enable Twitter users to tag their tweets for online use.

As a conventional utterance, *"#TweetComme"* ("#TweetLike") creates an overassertion inside the discourse. The latter is host to gender considerations when phrases such as "*"#TweetCommeUneFille"* ("#TweetLikeAGirl") or *"#TweetCommeUnMec"* ("#TweetLikeABoy") are produced. *"Fille"* and *"Mec"* work as the archetypal pair. Such a use of categorising keywords fits into a binary drift rendering either an auto-representation of one's own gender, or a hetero-representation of the other gender by the enunciating subject. By such a statement, the subject makes a prototype model of general (or virtual) scope out of *une fille* or *un mec*.

We will see how by over-asserting one's tweet with *"#TweetComme"*, one works actively to make the image of a stereotyped speaker –*fille* (a girl) or *mec* (a boy, a guy, a dude) likely to produce or to have produced said over-asserted speech– surface into the other Twitter users' minds. General tendencies will then be extracted to show how uttered clichés are associated to paradigms that are presented as gender-dependent: sports, alcohol, sex, etc. A sense of community-building activity comes to sight with Twitter as a social media becoming a hub from which are propagated collections of discourses which reception is oriented by the categorizing selection made through the hashtag.

2 About the corpus: why *"#TweetCommeUnMec"* and *"#TweetCommeUneFille"*?

By way of overture, it is important to list the parameters that allow defining the outline of the environment in which the studied speeches were produced. It is after all their social and technological bosom in many ways, with all implied constraints and potentialities.

2.1 Twitter: a micro-blogging platform where discourse is embedded in a maximum of 140 characters

Twitter is a micro-blogging platform where discourse is embedded in a maximum of 140 characters. It was created in 2006. Its core principle is for users to answer the question *What are you doing?* Just like any other speech, tweets host their author's discursive ethos: one's self-image painted in the way in which one intends to be seen. In 2013, Longhi outlined that said discursive ethos could be echoed by a *technodiscursive* ethos that can be produced y the means of the platform's technical jargon. Longhi also suggested a third dimension for self-image producing: the *pre-technodiscursive* ethos. The latter is produced when background elements (*"des arrière plans"*) are embedded into speech with the use of *prédiscours* that expand various elements (logical and semantic) that are preliminary to the utterance. We use the French word *prédiscours* as defined by Marie-Anne Paveau in her 2006 book *Les Prédiscours. Sens, mémoire, cognition.* For clarity purposes, we will not yet try to translate the word so soon in this article, as this endeavour would most probably interfere with previously attested terminology in academic literature in English. In this case, we are referring to pre-established, well spread speeches that hold a cognitive operation of categorization approved by *virtually* all members of the linguistic community from which said speech is produced. In other words a commonplace thought, a stereotype. Hashtags can be designed as tools to enact such pre-technodiscursive ethos in a given speech.

Also, as the social media it is, Twitter is often used to spread community-building injunctions, i.e. to urge members of a given community to collude in some way.

2.2 The hashtag: a tool for referencing tweets with keyWord-dependent tags

In Twitter's restrained enunciation context, the speaker's goal is to enhance his tweet's visibility by assigning a hashtag to it that will both tag it (or reference it) and categorise it on the WorlWide Web. As Paveau puts it, the hashtag is a "linguistic item, which function is essentially social, that allows ambient affiliation of users, techno-conversationallity and searchability of discourse" (*"forme langagière dont la fonction est essentiellement sociale, permettant l'affiliation diffuse* (ambient affiliation) *des usagers, la technoconversationnalité et l'investigabilité* (searchability) *du discours"*).

In addition to its referencing role, the hashtag embodies a cognitive operation that brings it epistemologically close to Sacks' *categorisation device* (2000). Auto-representation of one's own gender (ex.: *"C'est sur quelle chaîne la cérémonie des Ballons d'or please ? Ils sont tellement beaux les footballeurs en costards #TweetCommeUneFille"* ("On which channel is the Ballons d'or ceremony broadcasted please? Soccer players in tux' are so handsome #TweetLikeAGirl") – female twitto), hetero-representation of the other gender by the enunciating subject (ex.: *"C'est trop nul, ce soir les garçons vont parler que de foot #tweetcommeunefille"* ("Such a shame, tonight boys will only talk about soccer #TweetLikeAGirl") – male twitto).

2.3 *#TweetCommeUnMec* vs. *#TweetCommeUneFille*: an archetypal couple

The reason why the *fille/mec* opposition was selected for this study is it appears to be the most representative occurrence of the discursive activity of gender-discriminating categori-

sation in the french-speaking community. In the over-assertion activity materialised in the technomorpheme "*#TweetComme* + male categorising item", the tag *"mec"* (dude, lad) was most productive, more than *"garçon"* (boy) or *"homme"* (man). The same goes for the female categorising item *"fille"*, more often used than *"femme"* (woman), *"nana"* (chick) or *"meuf"* (very common *verlan* slang version of *femme*).

This shows users' preference for the opposition *fille/mec* to embody gender-discriminating categorisation.

In a more semantic view, the pairing of *fille* with *mec* operates a semantic setting of both words that gives rise to quite a precise, narrow referential range. The etymological dictionary of French language *Le Robert historique de la langue française* rebuts any historical link between *mec* and *mac* (short for *maquereau* (mackerel, "pimp")), or with *mag*, that would have linked it to the Latin mega meaning large. Le Robert describes the origins of *mec* as "obscure". Nevertheless, it is in the 19th century that *mec* made its way into French slang (*argot*). First meaning "a strenuous, vigorous, leader of a man", it eventually referred to any male individual. Later on, *mec* was defined by opposition to *nana* (chick), and its actual synonyms are *gars*, type or *homme* (lad, dude, fellow, chap, bloke, man or individual): a broad sense, but with a pejorative twist. As for *fille*, when paired up with *garçon* (boy), both *fille* and *garçon* are then to understand in their general sense of boy and girl. But paired up with *mec*, *fille* takes another colour now holding the sense of prostitute in potentiality.

3 A few words on our theoretical background

Here we will try to give a brief outline of what we mean by *stereotype* and perhaps even more about *stereotyping* as a social activity embedded into language use.

3.1 The *praxis*

The theoretical core of our approach is Praxematics. It is in *Introduction à l'analyse textuelle* by Lafont and Gardès-Madray, in 1976, that the word Praxematics (*"la praxématique"*, in French) first arose. It names a branch stemming from Discourse Analysis, in the field of Language Sciences. Praxematical discourse analysis was first developed in the late 1970's, early 1980's by Lafont, Gardès-Madray and Siblot (*Le travail et la langue*, Lafont, 1978 ; *Pratiques praxématiques*, Lafont, Gardès-Madray, Siblot, 1983), and is today mainly expanded through the journal *Cahiers de praxématique* and in the works of the laboratory *Praxiling* at Paul Valéry University in Montpellier (Détrie et al., 2001).

This sprout in discourse analysis focuses its attention on natural corpora; i e. naturally produced discourses, spontaneously uttered prior to investigation, rather than induced speech or selected fragments. It puts the subject's experience of the world and his relationship with reality at the basis of the linguistic representation of the world. Praxematics developed a conceptual tool to better depict its model for sense production in discourse: the praxeme. (Lafont, Gardès-Madray 1976/1989, p. 99, *in* Détrie et al. 2001, p. 263) *"Nous appellerons praxèmes les outils de praxis linguistique qui permettent le repérage de l'analyse du réel objectif par l'homme et spécialement le repérage des autres praxis."*

This *praxis*, borrowed from Philosophy and strongly influenced by Phenomenology, represents the cognitive processes that are nourished by past experiences –both perceptive and practical–, and that are motivated by intention-bound relations, and that result in sense (Détrie et al., 2001, p. 266, translated by us). In Détrie's words (2001), the praxeme is *"[l']unité lexicale [prise] au moment de son actualisation discursive, mais aussi en tant qu'outil linguistique porteur de potentialités signifiantes (capitalisées en langue) qui résultent (et reflètent) des savoirs acquis sur l'univers extralinguistique grâce aux praxis humaines."*

Moreover, and as early as the 1970's, this cluster of linguistic practices made it a major point that the context of production be given an essential role in all analysis.

It also borrowed to Gustave Guillaume a conceptual tool that allows it to look at the processes at stake in meaning production in language. Guillaume referred to said concept as the process of *actualisation*. In this new context, the word refers to the crossing from the *to be said* stage (*"l'à-dire"*) to the *saying* stage (*"le dire"*) and finally to the *said* stage (*"le dit"*). We are talking here of the changes in nature that occur when what the subject intends to say goes from the state of pre-language magma through the enunciation operation and to a speech-product. In other words, actualisation means the crossing from language (*"la langue"*) to speech (*"la parole"*).

This contribution allows us to clearly turn away from the diktat of interpretation, and decisively focus our attention on the emergence of meaning.

3.2 The dialectic of the same and the other

The Dialectic relationship between *Same* and *Other* hinges on two philosophical concepts, the *Idem* and the *Alliud*, that follow from the binary sorting that operates in the awareness process at stake as the subject partitions the world in order to categorise it more sharply (*"[le] tri binaire qui s'opère au niveau du découpage du réel"*) (Détrie, Masson, Verine, 1998 : 44-45). That this very movement between Same and Other is reflective of the construction of self-identity — refered to as the *Self* (or *Ipse*) – also makes it an access-point to the speaker's subjectivity.

To Jeanne-Marie Barbéris (1998 : 30), this linkage between *Ipse*, *Idem* and *Alliud* (Self, Same and Other) is as essential anthropologically as it is in terms of systemic representations.

In other words, if the analysis is about the to be said stage, the choosing of the right words, then the speaker's lexical selection is made through the generic cognitive operation that says x is not y, and y is not x: inclusion of Same, exclusion of Other.

3.3 About dialogism in nomination

The words we use are all second hand. Others who have proceeded to their own semantic tunings have already said them all. Those tunings themselves have sometimes been endorsed by the linguistic community (French-speaking in our case). We merely recycle others' discourses for the imperious purpose of being understood. This is the dialogism of nomination. Past uses of words and phrases have each time hosted their speaker's motivations. Those past uses eventually sediment their meaning programs. Those meaning programs then become available in the language in question, capitalised for other speakers' later use.

3.4 Ready-made in language

The image of a general movement of capitalisation in language is probably what best describes how ready-made can be found embedded in speech. French Discourse Analysis calls this *le préconstruit*. We will venture to use the English word *preconceived* for this article. As shown, what is preconceived in terms of meaning production calls on what has been said, or prior discourses. From now on, we will use the phrase *prior discourses* or virtual prior discourses also as a transposition of the French *prédiscours* (see 2.1.). In language, ready-made works like a massive volume of layers and layers of settled meaning programs that were once hosted by lexicon. Albeit settled, this volume is fluid and shifting: some usages disappear, some develop, all are bound to evolve. The *preconceived* can thus be seen as both the entry and the exit point of lexicalisation processes.

3.5 Lexicalisation

In his use of words, the speaker gives them a patina that, after being put through other discourses of other speakers, will potentially disseminate in all the linguistic community. This is lexicalisation. The *preconceived* weighs on the discourse, and the latter, by lexicalising by chunks (phrases), enriches and yet trends the former.

3.6 Folk wisdom

The meaning programs that are under sedimentation into the layered volume of the preconceived hold the precursors for what is referred to as *folk wisdom*, *conventional phrases*, *buzzwords*, *clichés*, *collective unthinking*, *collective psyche*, *stereotypes*, ... They stay in use as long as they are contained in the *preconceived* layers.

This can happen locally, at the interaction level: the *preconceived* then calls upon the interlocutors' conversational history. At the level of the linguistic community, the *preconceived* calls upon collective history.

3.7 Praxical patterns (or praxical schemes)

In order to use the praxematical terminology earlier presented in this paper, let us simply point out that this *preconceived* layering are what we call praxical patterns (*"schémas praxiques"* in Détrie's work). They are a sum of socio-experiential patterns that are shared and thus endorsed by a linguistic community, and that underpin their language's way of partitioning their world (see 3.2.).

3.8 Value-oriented stereotyping systems

This conceptual *millefeuille* that has been unwound notion after notion, leads to the idea that our chosen corpus of study conveys stereotypes that work in highly value-oriented systems.

This figure works as a model in the sense that it is meant to simplify general observations as well as to connect them to the conceptual framework earlier exposed.

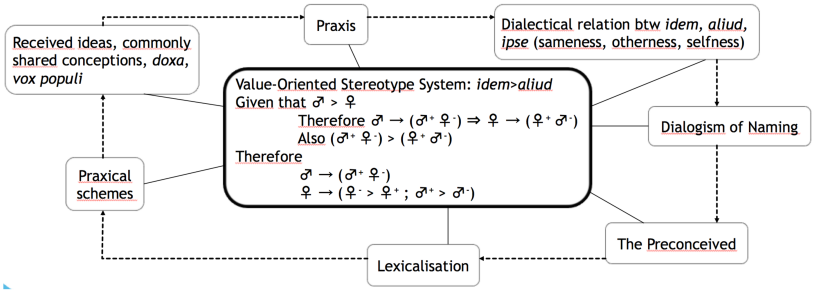


Fig. 1: Value-oriented stereotype systems.

Assuming that men dominate women, men produce positive representations of men and negative representations of women. In reaction, women produce positive representations of women and negative representations of men. Yet, as stated, men dominate women, therefore men-produced stereotypes dominate women-produced stereotypes; so there is no such thing as symmetry here. Men are impervious to women-produced stereotypes whereas women do tend to make men-produced stereotypes their own. As said, men produce positive representations of men and negative representations of women, women are influenced by ambivalent representations of themselves and positive representations of men: $♂ \rightarrow (♂^{+}♀^{-})$, $♀ \rightarrow (♀^{-} > ♀^{+}; ♂^{+} > ♂^{-})$.

4 Tendencies from our corpus

In this section, a few graphs show general statistical tendencies out of our observations of 1571 tweets. As said earlier, tweets were collected using both hashtags *#TweetCommeUneFille* and *#TweetCommeUnMec*, applied to all tweets published since the birth of Twitter in 2006 and still available at the time of the study.

4.1 Sexuality and Relationship

It appears that Twitter users put *relationship* status and romance in the girls' dominion. To a lesser extent, motherhood is also in the girl area as is *feminism*. In the guys' dominion are *infidelity*, *boorishness* and *singlehood*.

The global stereotype at stake is that women thrive in relationships – with its corollary romance and motherhood–, whereas men do not care for such questions.

Regarding sexuality, men are most associated with intercourse and sexual attraction, then to *homophobia*, *flirting*, *penis size*, and *street harassment*; whereas girls are less represented by their sexual life.

The global stereotype brings up a hypersexual man and a woman who gives much less importance to her sexual life.

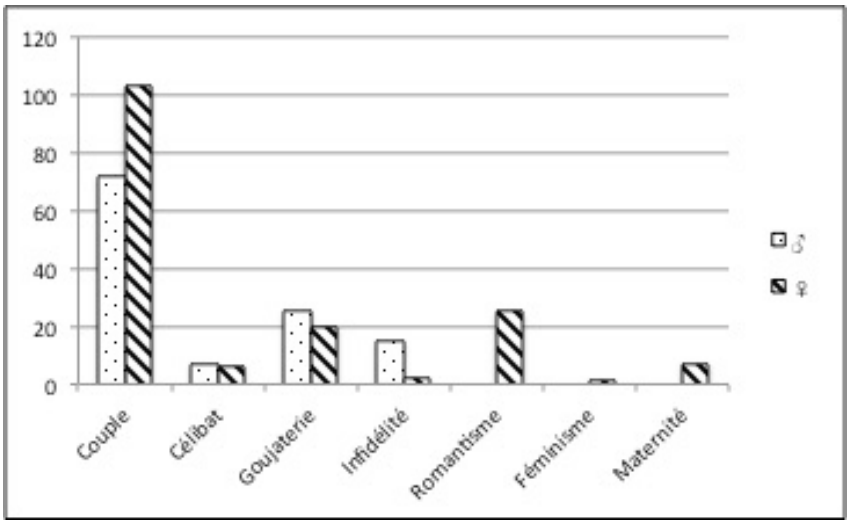


Fig. 2: Number of tweets linking gender to couple topics.

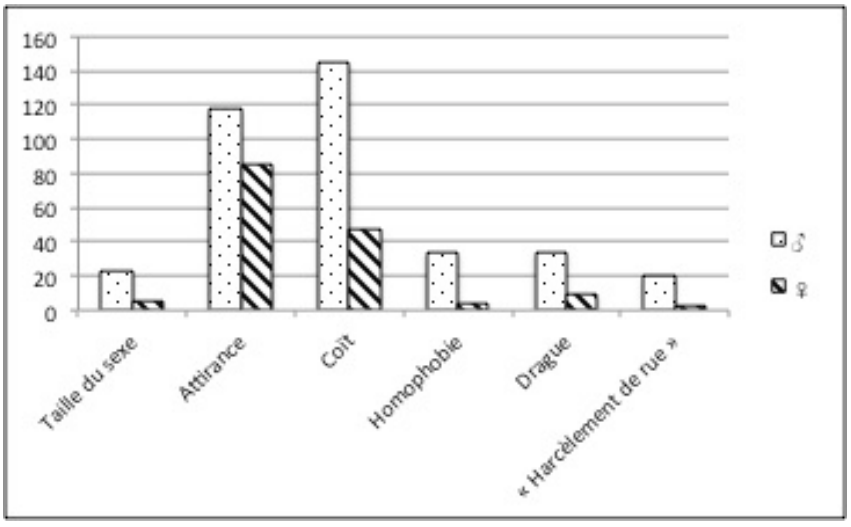


Fig. 3: Number of tweets linking gender to sexual topics.

4.2 Sports and video games

Men are massively represented as sport connoisseurs, whereas women are strongly depicted as ignorant in sport or even with a heinous point of view towards it. The stereotypical man is associated with soccer and workout. Women ignorance first concerns soccer, then rugby and then basketball. It is interesting to see that no tweet actually represents the stereotypical man as a sports ignorant or a sports hater.

The global stereotype at stake is that of a man who is a connoisseur, and a woman who either hates or knows nothing about sports. The sport that is most referred to –all genders

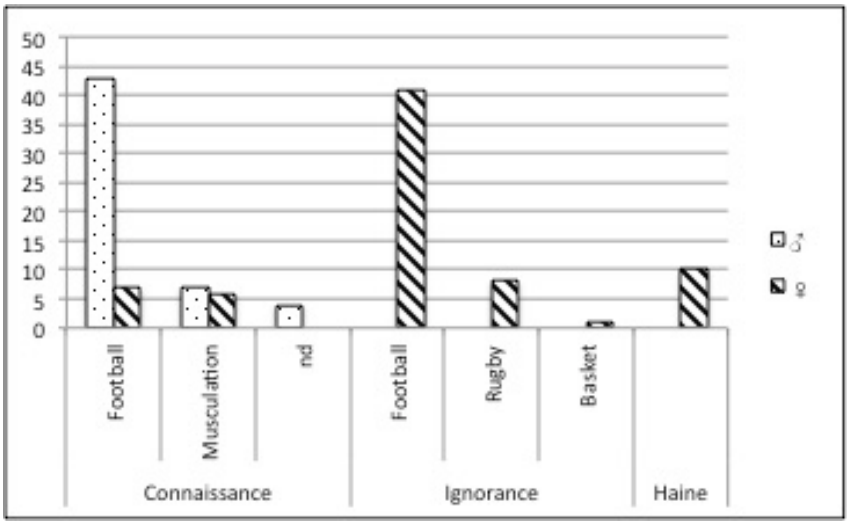


Fig. 4: Number of tweets linking gender to sports.

considered– is soccer (*football* in French). For men, the second sport is workout, and for women it is rugby (or the ignorance of rugby!).

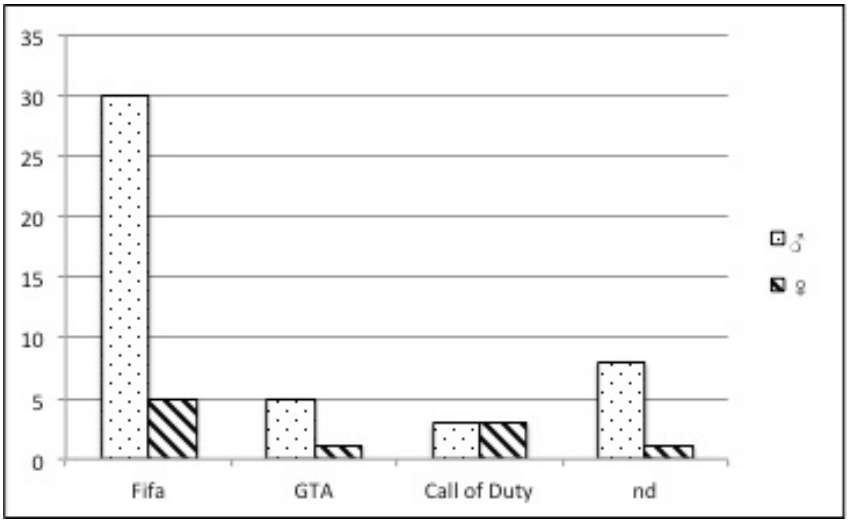


Fig. 5: Number of tweets linking gender to video games. (Fifa :soccer game, Grand Theft Auto: road violence game, Call of Duty: war game)

Men are more often associated with video games than women. It is very much the case regarding the *Fifa* game (soccer), a little less regarding videogames in general (*Undetermined*), and even less for *GTA* (road violence game). *Call of Duty* –a worldwide best selling video

game– is equally mentioned for men and women. However, the last two games do not count utterances enough to be statistically interpretable.

4.3 Culture

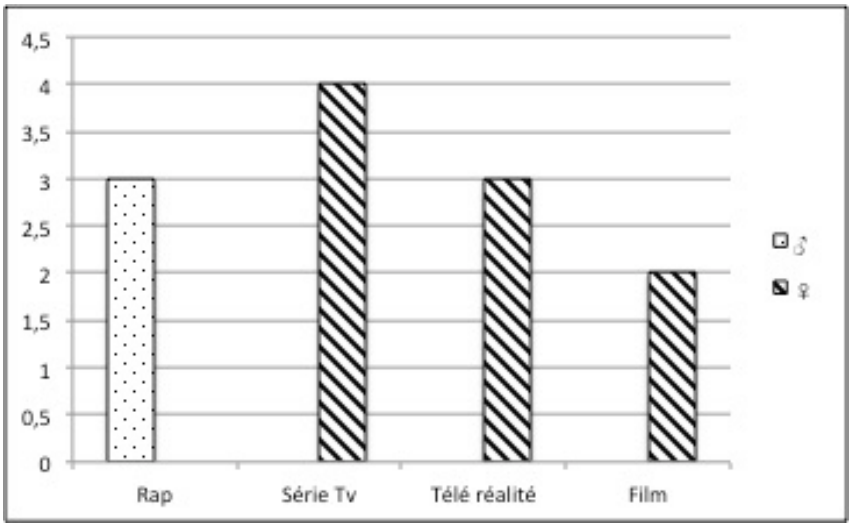


Fig. 6: Number of tweets linking gender to cultural products.

This time, a clean cut separates the girls from the guys: according to French-speaking Twitter users, the stereotypical man listens to rap, and the stereotypical woman enjoys TV series, reality shows and movies. No tweet links women to rap, and no tweet links men to TV series, reality shows or movies. Those are the only four cultural products represented in our corpus.

4.4 Avenues for Reflexion

Overall, among 1571 tweets, 1443 were posted by females, 773 by males, and 26 by Twitter users whose gender was undetermined.

We had 762 *#TweetLikeAMan* and 809 *#TweetLikeAGirl*, thus quite a well-balanced ratio.

Men were more likely to use the hashtag *#TweetLikeAGirl*: 706/67. Girls have a more balanced ratio: 771/672, with a slight tendency towards autostereotyping.

It is very difficult to say if a topic is linked to a gender in a negative or a positive way. For instance, men' infidelity is sometimes valued, sometimes sentenced. Irony and sarcasm are frequent and very easily misconstrued, especially in written speech. It is ever more true with written discourses that are so often decontextualised. This is why it was not possible for us to consider statistically studying positive and negative values linked with gender-oriented stereotypes in this corpus.

Nevertheless, it is interesting to see how topics related to a girl's tweet are far more diverse than those related to a guy's, far more homogenous.

5 *#TweetLikeAMan* versus *#TweetLikeAGirl*: highlights from the corpus and discussion

5.1 Commonplaces and gender-dependent topics

As we were reading for the first time the whole tweet collection, we noticed that the Twitter users repeatedly used (and thus endorsed) many widely shared stereotypes. We mostly found stereotypical linkage between topics and a given gender. But prior to that, let us take a look on two typical examples

> "ahah SEX ahah FOOT ahah TROP BONNE ahah BASKET ahah ZLAATAAAN ahah TOUTES DES PUTES ahah BUUUUT ! #TweetCommeUnMec" (♀)
> "#TweetCommeUnMec elle est tellement grosse que si je te la pose sur la tête sa te coupe la croissance." (♀)

By tweeting as *a guy would tweet*, these female Twitter users link those three topics –sport, sexuality and penis size– to the male gender. Given the humoristic character of both tweets, we are to expect both Twitter users to have posted them with the aim of being retweeted, in order to increase the reputation of their profile. Other tweets, while not specifying so clearly the topic they are exploring, also make use of humour by exploiting praxical patterns in a creative way:

> "Ma copine rêve que je l'emmène en croisière, alors je la mène en bateau #TweetCommeUnMec" (♀)

This last tweet would enter the *Boorishness* category in our statistical study.

> "Il était doux *___* c'était parfait. #TweetCommeUneFille Je l'ai ken d'une force ! Elle a pas assumé. Elle boîte depuis. #TweetCommeUnMec" (♂)

This tweet is symptomatic of the use of the selected hashtags: *#TweetLikeAMan* and *#TweetLikeAGirl*. This highlights the way in which those hashtags can be used to give off the impression that the user is an *expert* in the area of what differentiates men from women.

When a female user wants to speak crudely of sexuality, she is likely to use the hashtag *#TweetLikeAGuy*:

> "Le fréro delavega avec la veste la, bah jle déboite . #TweetCommeUnMec #The Voice" (♀)

By doing so, not only does she link crudeness and sex-talk to men, but she also shows how uncomfortable it is for a woman to talk this way.

The same goes for sports. A woman who is knowledgeable in soccer and tweets about it is likely to feel the need to say she is tweeting like a boy:

> "Tranquil contre l'Ukraine, Rio est déjà à nous les gars #TweetCommeUnMec" (♀)

> "Demain qui est chaud pour un foot ? #TweetCommeUnMec" (♀)

And if men are gossiping:

> "@Simonn_C jtexplique demain gros mdr ca nous occupera en maths #TweetCommeUneFille" (♂)

5.2 Hetero- Versus Auto-Representation

In tweets, stereotyping acts are either made in auto-representation or in hetero-representation. The cases we chose to put forward are those that involve the womaniser figure. One tweet is male-produced while the other is female-produced:

> "Je t'aime mon cœur :$ Envoyer à vannessa, Mélissa, Caroline, Leila, Lisa, Lucie, Emma.. #TweetCommeUnMec" (♀)

> "DemJe t'aime doudou1 doudou2 doudou3 doudou4 #TweetCommeUnMec" (♂)

Whether the Twitter user is female or male, both tweets draw the womaniser as an archetypal representation of the loving man.

5.3 Retweets and confrontational conversations

Earlier, we talked about value-oriented stereotyping systems (see 3.8.). Sometimes, other users did detect a pejorative orientation in another user's tweet in our corpus. This led to retweets and commentaries that sometimes developed into confrontational conversations. Interestingly enough, the most angry reactions mostly came from male Twitter users.

> "Vous êtes chaud avec #TweetCommeUnMec. Je vais me mettre a faire des #TweetCommeUnePute ma méchanceté atteindra son paroxysme" (♂)

> "@____ : #TweetCommeUneMec elle a un boule de ouff mec elle est bonne à niquer ! » Aucun mec ne parle comme sa." (♂)

Reacting male Twitter users are deactivating the stereotype by saying this is one guy talking, not all guys. They are contesting the archetype.

> "Je ne sais pas qui vous fréquentez mais je n'ai jamais dit un seul de vos #TweetCommeUnMec" (♂)

> "Depuis tout à l'heure je vois des #TweetCommeUnMec Certain serait plutôt #TweetCommeUnKsos" (♂)

This time, not only is the archetype contested, it is replaced by *"ksos"* (from *"cas social"*): a socially challenged person. It is a recategorisation that is taking place here.

This confrontational reaction is almost exclusively posted by male Twitter users. Female contestation is rather passive in comparison:

> "Avec vos hastag « #TweetCommeUneMec » et « #TweetCommeUneFille » on en conclus que les gars sont vulgaires et les filles Hypocrite et Chieuse" (♀)

> "@MenoJmen : Comme J'suis fat sur cette photo #TweetCommeUneFille `pic.twitter.com/lqlvYKIT` » Salaaaaud ahahaa" (♀)

(Pastiched conflict.)

This illustrates the conclusions of Figure 1, that show how a dominated group tends to abide by the negative stereotypes society produces about it.

> "Mddr la meuf elle dit « les bg venez » alors qu'elle ressemble a rien mdr si elle veut des bg faut être un minimum potable #TweetCommeUnMec" (♀)

This tweet states the obligation for female to be good-looking according to male (stereotypical) desire. Nevertheless, by using the hashtag *#TweetLikeAMan* it also states male hegemony over women.

In a similar double-talk statement, this last tweet exposes woman servility, and yet one could argue that the (female) Twitter user is most probably speaking her own mind, and then trying to show she is taking some distance from her statement by marking it with this hashtag.

> "C'est M D R les filles qui se plaigent h24 sur fb/twitter prsq un mec l'a paqué.. T'avais qu'a être mieux pour lui hyn #TweetCommeUnMec :p" (♀)

5.4 Associated Hashtags

There is no doubt that *fille* and *mec* work like an archetypal pair thus making us depict an opposition between a servile archi-woman and a macho archi-man. Yet, although the former is often present in our corpus, the latter only appears once:

> "Pouuuaaaaah [Surnom] clash trop ! Vengeance #TweetCommeUnMec ma femme jlui dis fait moi un café elle le FAIT point #Machovaaa" (♀)

Other tweets do depict a misogynistic archi-man, but female endorsement of negative female stereotypes always surfaces in some way:

> "J'ai besoin d'une femme la sa devient chaud chez moi qui va cuisiner le menage tout sa faut jme pose la #TweetCommeUnMec" (♀)

> 'Ma femme doit savoir cuisiner. #TweetCommeUnMec La femme en question sait même pas faire des pâtes. (je fais partie de celles là LL)" (♀)

In this last tweet, the female Twitter user is sorry not to fulfill her male partner's cooking expectations.

The housewife figure also comes in other forms:

> "J'ai besoin d'une femme la sa devient chaud chez moi qui va cuisiner le menage tout sa faut jme pose la #TweetCommeUnMec" (♀)

We see that users do not feel the need to say the words misogynistic or macho. The opposition between *#TweetLikeAGuy* and *#TweetLikeAGirl*, after being used and sedimented by the community, is already host to misogynistic and macho views that precipitate into praxical patterns (see 3.7.). This means that the decontextualisation at stake (see 4.4.) comes less from a will to take some distance from the discourse ("This could have been said by *x*."), and more from an active generalisation (i.e. stereotyping) movement ("All men are misogynists.").

> "Une fille en jupe n'est pas une pute messieurs mais vous des daleux #TweetCommeUneFille #feministeoulaferme #safedanslarue" (♂)

Referring to the Same and Other dialectical pair (see 3.2.), one can see here that even if this man feels concerned by female harassment, his use of the hashtag *#TweetLikeAGirl* no matter how compassionate, says that gender oppression remains a girl's concern.

5.5 Metadiscursive tweets and gender supremacy

As we showed, categorisation by these hashtags can start conflicts among Twitter users. Moreover, we have shown how little general social problems are discussed. The opposition between girls and boys is often represented in gender-dependent topics, incompatible views on relationships, and jokes. These very last tweets show a male Twitter user antagonising other male users over their use of the hashtag *#TweetCommeUneFille* in rather a violent way:

> "«#TweetCommeUneFille» VOILA UN HASHTAG QUI PERMETTRA AUX MECS FRAGILES DE NOS #TL DE S'EXPRIMER A L'AISE !!!! BANDE DE BOLOSS SODOMISES" (♂)

This is a metadiscursive hashtag in the sense that it is discursively looping towards itself: "I am talking about the hashtag I am using."

> "Pour certains garçon Le hashtag #TweetCommeUneFille leurs permet de dire vraiment ce qu'ils pensent des garçons !!" (♂)

What both tweets say is guys shouldn't tweet like girls.

6 Conclusion

The study of tweets that are categorised by the hashtags *#TweetCommeUneFille* and *#TweetCommeUnMec* forces us to think over the question of decontexctualisation-recontextualisation of reported speech. The use of those hashtags and the conventional utterances *#TweetComme* does mean an activity of over-assertion for the user is clearly saying "I am tweeting" ("and I am tweeting that I am tweeting"). *Fille* and *garçon* add to this a stereotyping act directed towards the gender opposition between male and female. This generalising movement of thought paired with this over-assertion and produced in a decontextualised mode has been qualified in 2010 as *unsubstanciated reported discourse* by Bertrand Verine (*"discours rapporté non avéré"*). Verine later said (2011): "la situation d'énonciation enchâssée est donnée pour effective et définie, mais [...] l'acte d'énonciation rapporté est représenté comme seulement plausible".

Twitter users use the preconceived -–or, as Authier-Revuz puts it "an inter-speaker dialogism of grater range" ("*un dialogisme interlocutif large*") (1984)— when they specify that they have chosen words that have been already used (by a specific gender). This is how they end up producing, endorsing and spreading gender-dependent stereotypes. This extreme overlapping of gender-dependent topics is a fundamental pillar of stereotypical structuration. The allusions to words from a virtual speaker compel the interlocutor because the speech needs to be understood to be powerful (KomurThilloy, 2010) despite the dissolution of one speaker into another (Authier-Revuez, 1984).

Bibliography

[1] Authier-Revuz J. (1984). Hétérogénéité(s) discursives. *Langage*, 73, 98-111.

[2] Authier-Revuz J. (1995). *Ces mots qui ne vont pas de soi. Boucles réflexives et non-coïncidence du dire, Tome 1.* Paris : Larousse.

[3] Barbéris J.-M., (1998), Identité, ipséité dans la deixis spatiale: ici et là, deux appréhensions concurrentes de l'espace?, *L'information grammaticale,* 77, 28-32.

[4] Détrie C., Masson M., Verine B., (1998), *Pratiques textuelles*, Montpellier: Presses de l'Université Paul Valéry.

[5] Détrie C. Siblot P., Verine B. (2001). *Termes et concepts pour l'analyse du discours. Une approche praxématique.* Paris : Honoré Champion.

[6] Komur-Tilloy G. (2010). *Presse écrite et discours rapporté.* Paris : Orizons.

[7] Longhi J. (2013). Essai de caractérisation du tweet politique. *L'information grammaticale*, 136, 25-32.

[8] Paveau M.-A. (2013). Hashtag. *Technologies discursives*, [Carnet de recherches], [En ligne]

[9] Verine B. (2010), « Construire une connivence dans la disjonction : l'emploi extensif de tu et le discours autre non avéré dans les commentaires rugbystiques d'Herrero et Cazeneuve », Dans C. Détrie, B. Verine (dir.), *L'actualisation de l'intersubjectivité de la langue en discours* (p. 89-104), Limoges : Lambert-Lucas.

Managing your online identity through Twitter within business-to-business (B2B) organisations

Lucill J. Curtis

Essex Business School, University of Essex
Wivenhoe Park, Essex CO4 3SQ
UK
lcurti@essex.ac.uk

Abstract. The purpose of this research is to examine the day-to-day management of organisational Twitter sites from the perspective of UK marketing professionals, within five B2B case-study organisations. Specifically this study examines how marketing professionals as 'organisational storytellers' engage in a process to communicate the official organisational story to customers, while maintaining an element of their personal identities, due to the less formal and brief (140 character) nature of tweets. This results in a fusion of organisational and self-narratives through conversational dialogue, appearing as 'narrative public voice' within the tweets. Adopting an abductive, interpretivist research design, this study initially analysed the content of tweets from five B2B Twitter sites over a three-month period, utilising a corporate web identity framework (Elliott & Robinson, 2013). The tweet examination informed the qualitative interviews with 15 marketing professionals across the five case-study organisations. The emerging results suggest a purposeful unification of personal/organisational identities as more marketing professionals are tweeting in a conversational voice designed to increase accessibility and interactivity with customers. Despite this most of those interviewed consider Twitter to be an awareness raising communications channel in the B2B sector, rather than a means of building long-term relationships with customers.

Keywords: B2B, corporate web identity, narrative public voice

1 Introduction: Aims

This study will provide a better understanding of the opportunities, challenges and barriers to utilising Twitter within business-to-business (B2B) organisations. In order to evaluate the potential of Twitter it is important to better understand the role, identity and behavioural traits of the 'organisational storyteller(s)', which for the purposes of this study are identified as marketing communications professionals working for B2B organisations. This is because it is the organisational storyteller who constructs the narrative that is projected via Twitter and which becomes the 'organisational story' and 'public voice' of the organisation. In the current literature the role of employees as online network facilitators is largely overlooked (Aula, 2010), so a better understanding of their role should also make an influential contribution to future marketing communications strategies, encompassing corporate identity.

By drawing upon the 'lived experiences' of the marketing communications professional(s) as the organisational storyteller within the virtual space, an opportunity arises to see how communication is developed and maintained in organisational Twitter feeds by those working day to day within the sector "in their own words and from their own perspective" (Tyler, 2011: 1483). This has been hitherto missing in the emerging literature relating to Twitter and online identity, which primarily reviews the experience from the perspective of the customer, rather than the 'organisational storyteller', within the Business to Consumer (B2C) environment, as opposed to the Business to Business (B2B) context, (Cover, 2012, van Dijck, 2013, Bullingham and Vasconcelos, 2013).

2 Constructing identity – can it ever be unified into a single narrative voice?

Social psychology literature (Harre, 1998) has described identity as predominantly singular and static, only liable to alteration gradually over many years. However, this theoretical approach has been disparaged in favour of identity as socially formed and open to change depending on the societal situation (Tajfel and Turner, 1979, Gergen, 1991). According to Karreman and Alvesson (2001: 63) "identity in contemporary social science… is seen as multiple, fragmented, processual and situational, rather than coherent, fixed and stable." Rosenberg (1997) draws upon James's (1890: 294) concept of the "empirical self" – where a person "has as many social selves as there are individuals who recognise him," to describe multiplicity. Based on social psychological practice this meaning derives from the suggestion that a multiplicity of selves has "a variety of conceptual partitioning's of self" (Rosenberg, 1997:23). Sassen (2000:11) explains this concept in relation to spaces for identity construction:

> "*Each one of us has multiple identities and multiple sites for action; worker and workplace are but one type… community, household cultural practices, our bodies – these are all sites for identity and for action.*" Sassen (2000: 11, as cited in Valle and Torres, 2000).

Arguably, this view reiterates the multiplicity of individual identities, espoused in modern society. Notwithstanding some of these identity constructs may co-exist in conflict, such as professional, (work roles) domestic (wife/husband, mother/father, friend) and group identities. This may be considered problematic. In fact Lawler (2008:3) suggests the main risk to stability and the consistency of self, is in what she describes as its "internal homogeneity," – the understanding that the consistency of one identity is simply not sustainable for anyone. Perhaps though there may be situations for professional identities where consistency and a singular personification of self may be desirable. Perhaps this would require one or more aspects of an individual's identity to reach a point of conscious compromise, or for one or more identity constructs to be omitted in favour of the 'dominant other'. For example the young female social worker, recently enjoying the birth of her first child, asked to remove a young child from their family due to suspected abuse would have to distance herself from her identity as a mother in order to comply with professional bureaucratic process. There may

though be professional roles that require less distinction between professional and personal identities. For example individuals running their own businesses, working freelance or within industries where their values are reflected within organisational values, may be able to unify elements of their multiple selves more easily. Scripted roles such as those people working in job centres and within customer service centres who are trained to respond in a united, consistent manner may also feel less tension between aspects of their identities (Hopfl, 2002). There is also the possibility individuals may find a way of managing their multiple selves into a single voice. Either they are performing their identity and therefore they speak with the unified voice of an actor, or they find a way to combine the qualities and ideals from their personal identity construct(s) with their professional identity. This may be ratified individually or through social interaction. Arguably this is where the online space may provide a less rigid, controlled place, where personal and professional identity may be seen to more comfortably co-exist or potentially unify. This potential co-joining of personal and professional identities online provides a slightly different perspective on more traditional identity work theories and one that differs from the earlier theories on online communication, where identity is seen almost exclusively in terms of its multiplicity, as is outlined in the next section.

3 Existing online identity theories as multiple constructs

It may be considered that each participant in Twitter is trying to condense what could be termed his or her 'multiple selves'. However, many contemporary social theorists, Hayles (1999) and Braidotti (2013) in particular, have enquired whether our basic understandings of identity are changing as communication flows through computer mediated networks. Rather, are the Twitter generation more comfortable with multiple, even contradictory performativities, not resolved into a single, narratively stabilised identity. Within her seminal work 'Life on the Screen' (1995) Sherry Turkle explains the appeal of creating multiple identities, within virtual spaces, known as Multi-User Dungeons (MUDs). These sites, originating from "role-playing games such as Dungeons and Dragons let people take on fictional personae and play out complex adventures," (Turkle, 1995: 180). From the interviews included within the book, it is suggested young people overcoming shyness, loneliness and desiring escapism from the routine nature of their everyday lives, predominantly accessed these sites during the 1990's. It gave them the opportunity to 'live out' multiple existences, distinct from their own offline identities. She explains the nature of the communication within the online environment, as it was then:

> "*In the story of constructing identity in the culture of simulation, experiences on the Internet figure prominently, but these experiences can only be understood as part of a larger cultural context. That context is the story of the eroding boundaries between the real and the virtual, the animate and the inanimate, the unitary and the multiple self...*"(Turkle, 1995:10).

The creation of multiple identities in this scenario gave the participants freedom of expression without fear of reprisal, as their online identities during this period were predominantly

anonymous. For a link to be found between their identities online and their actual identity was very unlikely. Usually it was their choice when or if to reveal their 'true selves'. However, this is not the case today and most online identities are clearly visible. On a Twitter account for example an organisation uses extensive branding to actively promote their identity online. Organisations also often include their 'real' name and a photograph, usually including a logo to visually announce who they are.

This is a contemporary phenomenon since the Internet's only enabled virtual communication of this kind recently. However, earlier communication methods may be described as using a virtual medium, such as the written letter, semaphore, beacon-lighting or phone calls. They are considered virtual in nature as the communicators cannot physically see each other and so the recipients are reliant on an exchange of messages. Therefore it is important to recognise the newness of virtuality enabled by the Internet, but also that it has a precedent in other forms of absent-present communication.

What is evident from existing interpretations and applications of narrative identity constructs are the problems associated with the current definitions of identity both within the physical and specifically the virtual environment. McPherson (2000) for example suggests the time has come to move away from considering identity as play-acting online, instead an analysis of politics and participation is urged to form an awareness of authentic virtual world interactions. Therefore, an understanding as to how the organisations' identity is communicated through the organisational storyteller is intended, as well as an awareness of how this impacts on the self-identity of the individual employee(s). Initially an analysis of corporate identity constructs is discussed below, so as to consider the impact of organisational identity on the self identity of the organisational storytellers.

4 Constructing identity – can it ever be unified into a single narrative voice?

The study of corporate or business identity has focussed on those aspects of an organisation that make it distinct, and therefore identifiable to customers and employees alike (Balmer, 1998, 2001, 2008, Christensen and Askegaard, 2001). This is aligned with the concept: "*... that the effective management of an organisation's identity results in the acquisition of a favourable corporate image and, over time, of a favourable corporate reputation, which leads an organisation's key stakeholders and stakeholder groups to be favourably disposed toward it.*" As an aspirational concept corporate identity has been explored principally through formal communications literature and policies (Balmer and Wilson, 1998). Recently, however marketing (Abdullah, Nordin and Aziz, 2013) and organisational (Sillince and Brown, 2009) scholars have started to investigate the processes and challenges of maintaining a consistent and appealing corporate identity through online communications, specifically corporate websites. Abdullah, Nordin and Aziz, (2013) approach the study of corporate identity through an analysis of the mission and vision statements on the websites of 860 Malaysian and 642 Singaporean companies, utilising Aaker's (1986) Big Five brand personality dimensions (sincerity, excitement, competence, sophistication and ruggedness). The paper concluded

however that both sets of organisations were poor at positioning themselves distinctly through brand personality dimensions online, potentially leaving them at risk of failing to achieve competitive advantage and of maintaining their reputation. Nonetheless, little is known currently about the routine, day-to-day lived experiences of marketing professionals that support, produce and represent the brand.

Within this study the brand as constructed and embodied within the organisational story interpreted and communicated through the marketing professional's voice within the tweets is considered. Therefore an analysis of the day-to-day interaction and communication of brands, by producers and customers within an online Twitter environment is sought. In this sense an investigation of how corporate and self-identity unite within communications practice and emerge "*through self-aware reflections about whom one is and through everyday practices of doing work*" (Wieland, 2010) will be clarified. Such a sense of consistent unification may for example be achieved through the adoption and reconstruction of brand values. For example recent papers more explicitly review the relationship between brands and employee identity within contemporary organisations (Brannan, Parsons and Priola, 2015; Vasquez, Sergi and Cordelier, 2013). This paper will expand upon both papers to explore how marketing professionals engage in identity work when communicating with customers, through conversational discourse within the tweets, including the process of reconstructing and reaffirming the organisation's brand, as well as their own narrative identity as discussed below.

5 Narrative identity

Narrative is described by Webster (1966:1,503) as "discourse... designed to represent a connected succession of happenings." This reading of narrative alludes to the importance of vocalising experiences as a means of making sense of them (Weick, 1995). Although as early studies show narrative is also understood as written text. For example the study of narrative text is located in Aristotle's Poetics and initial hermeneutic investigations of the Bible and Koran (Czarniawska, 2004). However, it wasn't until the publication of Vladimir Propp's Morphology of the Folktale in 1928, which examined narrative structures within fairytales that it reached broader recognition as a method of textual analysis, principally. Its popularity grew as a means of communication and by the 1970's narratives were considered as ubiquitous, with an ability to adapt their shape and form to suit different social environments as Barthe's outlines below:

> "*Narrative is first and foremost a prodigious variety of genres... Able to be carried by articulated language, spoken or written, fixed or moving images, gestures, and the ordered mixture of all these substances; narrative is present in myth, legend, fable, tale, novella, epic, history, tragedy, drama, comedy, mime, painting... stained glass windows, cinema, comics, news item, conversation. Moreover, under this almost infinite diversity of forms, narrative is present in every age, in every place, in every society... it is simply there like life itself.*" Roland Barthes (1977: 79).

Interestingly this description suggests narrative is understood to be more than just the spoken and written word, it is present within images, gesticulations and theatrical performance. To show how individuals may be said to perform identity narratives I will discuss relevant theories (Goffman, 1959, Butler, 1990 and Boje, 1991) later in this section. The idea of narrative being visual as well as textual is a relatively recent conceptual development and is noteworthy here. In 1995 Julia Murray described narrative illustration as the "pictorial representation of or reference to one or more events that occur in a sequence of time and that bring about a change in the condition of at least one character" (Murray, 1995:17). Of those examples provided above by Barthes (1977) the stained glass window and painting suggest visual narratives may be viewed in isolation of text and still provide adequate meaning. However within film, comics and televised news reports, narrative and visual narratives are mutually complimentary, with one often helping to explain the other. To outline the challenges of understanding and producing narratives I will now explain the difference between a story and a narrative. This is important as this study examines how the organisational storyteller communicates the official organisational story through narrative.

McAdams and McLean (2013) describe narrative identity as "a person's internalised and evolving life story, integrating the reconstructed past and imagined future to provide life with some degree of unity and purpose." The idea of narrative identity as an evolving process is useful here in several ways. Firstly it suggests the incomplete ongoing and developing nature of narrative, which would work in an SMN environment, as the 'list' of tweets moves at a fast pace and is conversational in nature as outlined in the introduction to this chapter. The concept of the 'imagined future' also alludes to a certain degree of creative flexibility with the use of language. While the internalisation of the story also highlights something of an internal struggle possibly for the narrator who has to unify his or her multiple identity facets with elements of the organisation's identity through narrative public voice.

6 Introducing the new concept of 'Narrative Public Voice'

As outlined above there are various different competing demands on the organisational storyteller when they are communicating via Twitter, one element is the challenge of maintaining the consistency of one organisational identity narrative. This is what current marketing literature (Nandan, 2005) suggests the organisational storyteller should be striving for to secure brand consistency. One of the contributions to research this study offers is to develop a new theoretical construct focussing on the concept of a single unitary identity expressed through voice within the online environment. Voice in this context (specifically narrative public voice) is a different theoretical construct to previous interpretations of voice within organisational studies (Boje, 1995; Vickers, 2005 and Purcell, 2012), feminist theory (Belenky et al, 1997); and previous employee relations' papers (Budd, Gollan and Wilkinson, 2010). Voice is being investigated as the principle way for organisational storytellers to combine their self-narrative and the official organisational story. This enables the organisational storyteller to effectively 'tell' the official story, in a way that encourages them to retain elements of their self-identity, while also effectively communicating the organisational narrative in a way

that can be understood by customers. This model is being developed during the fieldwork process for this study and the findings are in the process of being analysed.

7 Method

7.1 Data collection

The first stage of the data collection included the collection of tweet data from Twitter sites of the five case-study organisations over a period of three months. The tweet data was then analysed for key themes, which determined the content of the questions for the next phase of the research. The second stage of data collection included qualitative semi-structured interviews with those responsible for organisational storytelling, including 15 marketing professionals, within each of the five case-study organisations. Once the interviews were transcribed the content as individual quotes, has also been grouped under the five core headings from Elliott and Robinson's (2013) framework. A summary of the characteristics for each of the organisations is included in Figure one below:

These organisations were chosen as they represent a diverse spread of B2B organisations offering an expansive range of products and services. This supports the validity of the data, as it provides an opportunity to explore the lived experiences of organisational storytellers in a variety of organisational contexts. In this sense each organisation is unique and may be treated as a stand-alone unit of analysis (Eisenhardt and Graebner, 2007).

8 Research design

8.1 New advances in qualitative research methods

The aim of this study is interpretive in nature, with a focus on the social constructivist paradigm, as it is focused on "how people craft their identities through interaction or how they weave 'narratives of self' in concert with others and out of the contextual resources within their reach," (Alvesson, Ashcraft and Thomas, 2008:8). A two-stage qualitative approach has been undertaken that focuses initially on a thematic narrative analysis of the tweets, followed by semi-structured interviews with organisational storytellers (marketing professionals responsible for writing tweets and managing the organisational Twitter site). Denzin and Lincoln (2005) summarise the essence of qualitative analysis as relevant to this study empathetically:

> "*Qualitative research is a situated activity that locates the observer in the world. It consists of a set of interpretive, material practices that make the world visible... attempting to make sense of or interpret, phenomena in terms of the meanings people bring to them.*" (Denzin and Lincoln (2005: 3).

Only a qualitative analysis, with its emphasis on reflexivity (Hammersley and Atkinson, 1995) has ensured an explicit emphasis on the detail of individual lived experiences of the organisational storytellers, especially as a lot of the activity, such as the twitter creation

Table 1: A summary of the characteristics for each of the B2B case study organisations.

Industry Sector	**No. of Employees**	B2B Customers	Markets	Primary Products/Services	Number of Tweets
Telecommunications	19,800 globally	Large corporate multi-national companies (MNCs), small, medium enterprises (SMEs) and start-up companies (entrepreneurs) including freelance workers	More than 170 countries globally	Broadband and internet support services, IT support and security, phones and phone line management.	1,437
Financial services	38,500 globally	Large corporate multi-national companies (MNCs), small, medium enterprises (SMEs) and start-up companies (entrepreneurs)	120 countries globally	Independent assurance, tax and advisory services	460
Healthcare	8,000 approx. in the UK	Local Government, specifically County Councils who are responsible for outsourcing healthcare contracts (state-funded elderly care homes) locally	UK only	Care homes, homecare and day care clubs	623
Business support	7,100 globally	Principally, SMEs, MNCs and, freelancers and entrepreneurs	2,000 business centres in 100 countries globally	Office space, meeting facilities, virtual offices, video conferencing and support services	382
Manufacturing	2,150 globally	Commissioners, Facilities Managers, Mechanical design engineers and building services experts	15 factories in 7 countries globally	Industrialised fans and ventilation systems	54

process, is intrinsic and thus not necessarily observable. Of particular importance here is a search for meaning through these experiences that sheds new light on emerging fields such as online identity through the observation and collection of organisational tweets. Interviews with the organisational storytellers, who create the tweets, has also encouraged an understanding of the process/role of the narrator and the practices they go through in order to create and respond to the content on organisational Twitter sites. This is pertinent as to date the emphasis within studies of Twitter has been on quantitative analysis (Krishnamurthy, Gill and Arlitt, 2008; Java et al, 2007). The regularity of features such as hashtags, retweets and followers lends itself to statistical analysis, arguably at the expense of a richer examination of understanding from the organisational storytellers themselves. Such insights from the storytellers will help to provide details of the processual nature of their 'identity work' (Watson, 2008) and the influences on their online 'identity construction' (Alvesson, Ashcraft and Thomas, 2008) as contained within the tweets and their interview narrative. To this extent this study proposes a deeper understanding of Twitter data that's different from simplifying similarities (Marwick, 2013). The adoption of a qualitative approach also allows a richer analysis of the storyteller's subjective interpretations of their situation through living story, evident within the tweets.

8.2 Collecting data from 'Living Story Spaces'

Given that this study has its origins in narrative tradition and is influenced specifically by Boje's concept of 'living story spaces' (Boje, 2011) it is appropriate that this framework is utilised to shape the methodological approach and to categorise organisational sensemaking, a key outcome from this data analysis. The constant fluidity of the online Twitter environment may be considered a place for 'networking in the unfolding present' and therefore a 'Living story space', worthy of further investigation as initially depicted within an adaptation of Boje's (2011) Storytelling Triad Model, Figure Two below:

Another important concept included in the triad model is 'antenarrative', which is considered by Boje (2004) as a 'bet' on the future of a past story, which has the potential to become active again and alter standard narrative constructs. As Boje (2011: 3) suggests antenarratives may appear as "fragments that seem to cling to other fragments, and form interesting patterns of assemblage relationships to context and one another." Given their brief, unordered, often, fragmented conversational nature tweets may be considered as a form of antenarrative interplaying in a living online story space. They may also be described as offering a link to historical narratives based on the organisational story, whilst interacting with developing living stories (Boje, 2011). In this regard such a living story space should also lend itself to the concept of a sensemaking narrative. As Brown, Stacey and Nandhakumar (2013: 1036) summarises:

> "*. . . theorists with an interest in sensemaking have argued that narrative is a primary cognitive instrument (Mink, 1978; 131; Polkinghorne, 1988) which constitutes the basic organising principle of human cognition*" (Boland and Tenkasi, 1995).

The importance of combining narrative and sensemaking within the same theoretical framework for analysis has lead to the utilisation and adaptation of Rosile et al's (2013: 3)

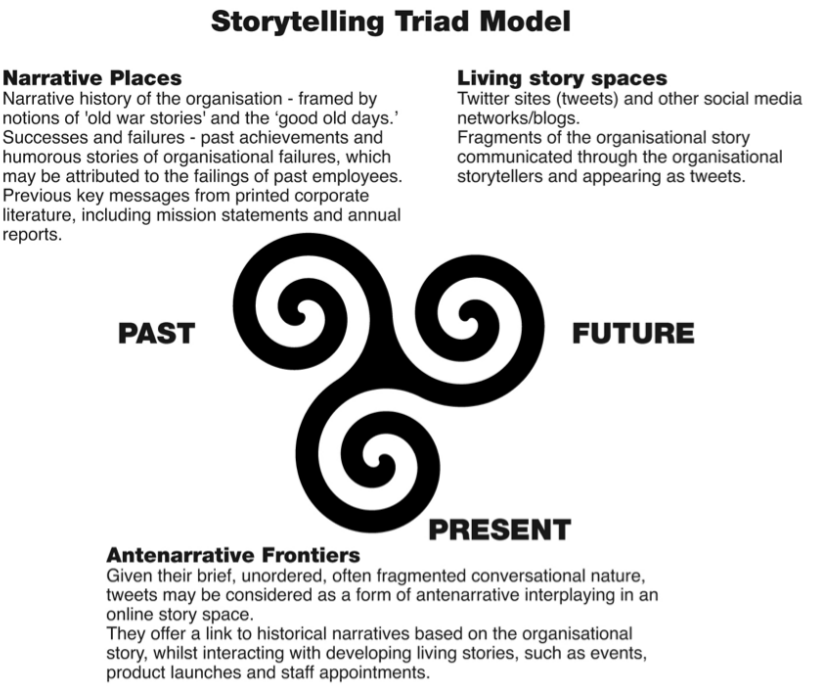


Fig. 1: Storytelling triad model – Adapted from Boje (2011).

storytelling diamond, as the overall means of shaping the methodological approach, which includes the tweet analysis and the qualitative interviews. Figure three, below provides an illustration of the storytelling diamond. Of particular interest is the 'living story' category and the arrows, which are the antenarrative processes moving in-between the paradigm (Rosile et al, 2013: 559).

Before application of the storytelling diamond is possible, Rosile et al asks researchers to consider their motives from epistemological, ontological and methodological perspectives as illustrated in Table 2 (Rosile et al, 2013).

Table 2: Epistemological, ontological and methodological perspectives (Rosile et al, 2013)

1	What is the epistemological value of storytelling? Is it a series of unique and generalisable accounts of reality, or is it a phenomenon that explains identity and rationality?
2	Does storytelling give a better understanding of an experienced truth of Being-in-the-world, or does it reflect a reinterpretation of lived experiences?
3	What methods motivate the researcher? Is the goal finding empirical evidence (both qualitative and quantitative) that leads to the testing hypotheses, or is it adding multiple perspectives so that a greater depth of understanding is possible?

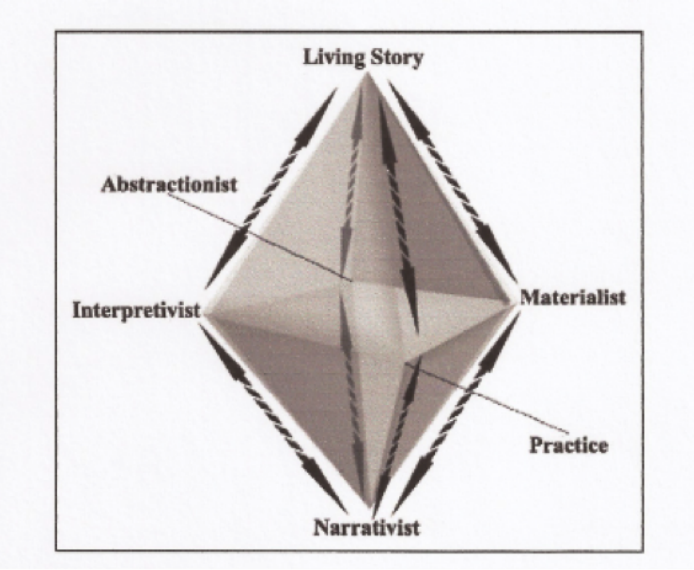


Fig. 2: The storytelling diamond (Rosile et al, 2013: 559).

Of particular relevance to this study is the second question, which relates to lived experiences. The combination of tweet analysis and qualitative interviews is intended to provide greater insights into the day-to-day experiences of marketing professionals within their role as organisational storytellers. Arguably the online dimension to these experiences provides a means of reinterpreting lived experiences, which is different to previous studies (Jacobi, Freund and Araujo, 2015) however the notion of "an experienced truth" (Rosile, 2013: 560) is also sought through the living story within the tweets and the narrative from the qualitative interviews. While Elliott and Robinson's (2013) web identity framework, discussed in the later sections (specifically in relation to the initial tweet analysis), where narrative and voice are initially linked to sensemaking.

Elements of the storytelling diamond model are highlighted throughout the rest of this paper to illustrate how it has shaped the data collection process, with links also to how it will inform the analysis. In order to make this work cohesively, Rosile et al (2013) suggests the researcher needs to be part of the process of living story, with an ability to recognise when antenarratives are interjecting into the present. Such a process of understanding is suggested through the analysis of the tweets by the researcher, but also through speaking with the organisational storytellers. This has supported understanding of the identity of the five case-study organisations and also their stories, past and present, so as to recognise the occurrence of antenarratives within the tweets.

As Rosile explains:

> "*If the audience with whom the scholar wishes to converse is in need of abstractions in order to value the findings, then the deep descriptions of living story should be turned into abstracted categories*" (Rosile et al, 2013: 570).

Such 'abstracted categories' of living story were identified from the tweet analysis utilising the broad categories as identified by Elliott and Robinson (2013) as part of their work making sense of websites and corporate identity thematics as detailed within the next section.

8.3 Understanding the context of the 'Living Story Space' - The collection of tweet data

The first stage of the data collection included the collection of tweet data from Twitter sites of the five case-study organisations over a period of three months. The tweet data was then analysed for key themes, which determined the content of the questions for the next phase of the research. The second stage of data collection included qualitative semi-structured interviews with those responsible for organisational storytelling, through the adoption of the narrativist paradigm (Rosile, 2013). This approach encouraged understanding as to how they make sense of their work/organisational experiences, within each of the case-study organisations.

An explanation of why these methods and case-study analysis are the most appropriate to meet the aims of this study and how they were undertaken, is outlined within the following sections.

9 Phase one – Antenarrative processes

9.1 Living Story Spaces analysis - Netnographic observations of Twitter sites from five case-study B2B organisations

Tweet data including text with links, images and video content was collected every three days for three months, from online Twitter sites (for each organisation), managed by the five B2B organisations. Every three days was a practical collection time for the tweets as the organisations wrote between 2-10 tweets on average most days. Therefore collecting every day would have meant a very large number of small files, whereas every three days provided fewer files with more data on each screen-shot. When collecting the tweet data a technique known as 'screen-grabbing' was used, which is similar to 'cutting and pasting' in Word documents. As illustrated below I was able to collect approximately 3-6 tweets for each organisation in one 'screen grab', Figure Six. However if images or video are included in the data capture there are slightly fewer tweets per 'screen grab'. This is because the images and video clip links take up a lot of space on the page as they are much larger than space required for 140 characters of text.

As Kozinets (2010: 56) explains, the observation of "*online community conversations and other internet discourse combines options that are both naturalistic and unobtrusive.*" Analysing tweet data through netnography allows the researcher to meet the principles of traditional ethnography, telling a believable, detailed and genuine narrative (Fetterman, 2010). Observing the tweets as they appeared in 'real-time' contributed to the sense of seeing a naturalistic narrative (living story) unfold, without the organisation or the followers of each Twitter site being conscious of the observation, as the researcher did not make any

Fig. 3: Example of tweet data 'screen grab' – the method of collecting tweet data during phase one. This Twitter page includes primarily text-based tweets, produced by Grant Thornton, [one of the five B2B case-study organisations], by the organisational storyteller over a four day period.

interventions on the site. This is important, as the researcher wants to observe the twitter site as if they are a 'customer' initially learning and making sense of the organisational identity

and that of the narrator, for the first time, through the type of language and images contained within the tweets. As a potential or pseudo customer there is the potential to observe what there is to glean in terms of data and clues about the identity of the organisation. The idea here is that the tweets highlight to customers those issues of most interest or significance to the organisation and therefore are where researchers should look for data/clues about how the organisation wishes to portray itself. As the content of the tweets was primarily informative, the messages are described as "closed netnographic texts" (Kozinets, 2010: 170).

10 Living story – Organisational and personal identity through story

The researcher views each tweet in detail while also trying to understand the broader themes and considerations relating to the organisation's and the organisational storyteller's identity. The process as Markham (2005:794) concurs was largely "*iterative; one's attention shifts alternately between close examination of texts to larger sensemaking frameworks.*" As Davies (2008) elaborates: "'*There are two primary ways in which ethnographic research on the internet can be contextualised. In the first place, attention should be given to the offline context in which internet usage occurs.*" In this sense the organisational setting where the tweets are written, as well as the identity of the organisational storyteller is of principal interest. This is because of varying levels of interaction by the customer accessing the organisation via online platforms (Sawhney, Verona and Prandelli, 2005).

Due to the volume and variety of tweets, this first phase of data collection provided a detailed understanding of the organisation's style of tweets (such as product promotion, new initiatives, events etc) their followers, hashtags used, retweets included and complaints experienced. The tweets also gave the first insight into sensemaking of the organisational narrative by the organisational storyteller – as encapsulated in the analysis of the narrative public voice. As Garton, Haythornthwaite & Wellman, (1995: 75, as cited in Kozinets, 2010: 50) explain, this type of analysis provides a means of understanding sequences of language formation and customer responses and silences (who speaks, who does not communicate and what is said proactively or in response to whom?):

> "*Social networking analysts seek to describe networks of relations as fully as possible, tease out the prominent patterns in such networks, trace the flow of information (and other resources) through them, and discover what effects these relations and networks have on people and organisations.*" (Garton, Haythornthwaite & Wellman, 1995:75).

The main purpose of this stage of data collection was to understand the information contained within the organisational tweets. As Brown, Stacey & Nandhakumar (2008) outlines the essence of narrative (and sensemaking) analysis is focused on three main elements, context, content and form. The contextual aspect of Twitter relates principally to who tells what to whom; the frequency of organisational tweets as opposed to re-tweets and the identity of the key spokespeople is therefore important. While content requires more specific analysis

of language – in terms of what is spoken and how frequently; and form relates to how it is said including the tone, words and the links and images used in conjunction with the text. An analysis of the content of the tweets initially collected as living story, was then used to identify the nature of the questions to be included within the interview scripts for the organisational storytellers.

11 Abductive interpretivist approach - Data analysis, interpretation and presentation

11.1 Using coding to identify themes from the tweet data

Following the identification of Rosile et al's (2013) storytelling diamond model as a lens with which to frame the two processes of data collection, the specific leitmotifs as originating from the tweet analysis are considered. An established qualitative coding model has been used to categorise the downloaded text, image and video files. Given the relatively small sample per organisation the researcher was able to start the process of interpretivist extrapolation of emergent living story themes into a draft framework, comprising slight modifications and additional aspects of the corporate website identity model. The main themes in the original framework for abstracting corporate web identity were: "mobility, accessibility, visuality, interactivity and customisation" (Elliott and Robinson, 2013: 276) and these are developed within the framework for analysis of the tweets below:

Mobility	What it is within the tweets that encourage movement/action? • types of revolving text • repeated tweets • links, moving the reader away from the central text
Accessibility	How customers make sense of the identity of the storyteller and the mass of information that is included • identification of the audience, who is the organisational storyteller speaking to? • complaints
Interactivity	The interplay between self and organisation • attempts via tweets to persuade customers to interact with the organisation/organisational storyteller positively • personal twitter feeds interjecting into organisational conversations • cross-over with followers
Visuality	Influence and occurrence of videos, images and hashtags. Content of images and how they contribute to organisational/organisational storyteller identity and narratives.
Customisation	Key messages and repeated phrases relating to the identity of the organisation and its story. Appearance of branding and consistency.

This framework has been utilised to structure the development of the abductive (Mantere and Ketokivi, 2013; Martela, 2011) coding process from the tweet data (Bryman and Bell, 2007). An example is included to illustrate how this process has been implemented to analyse the tweet data including the coding manually by hand as depicted below in Figure 4.

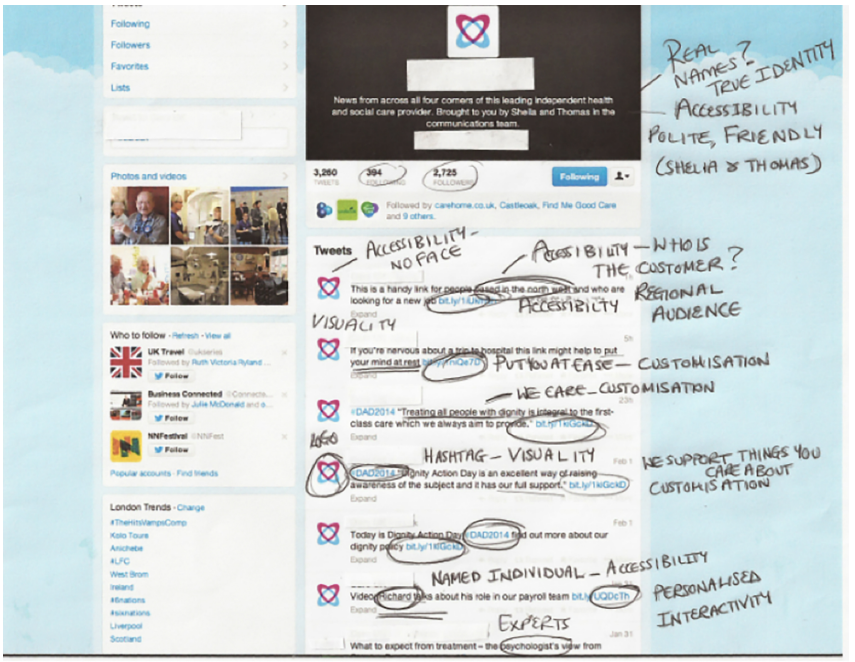


Fig. 4: This is an example of where the tweet data from one of the case study organisations has been manually coded using the headings and descriptions from the Elliott and Robinson, (2013) framework.

The next stage of the process was to use these headings (Elliot and Robinson, 2013) and their descriptions to develop a list of codes for each of the case study organisations (Kozinets, 2013), which led to the drafting of specific questions utilised within the interviews.

Arguably, the identity of the customer is inferred by the content, tone and style of the organisational tweet. Even if the tweets do not always promote a response it is suggested that each one is intended to engage existing and potential new followers. Engagement is encouraged by linking online content about the organisation, for example the links in tweets frequently refer the reader back to the company website. Tweets may also be originated in response to other online communities and help with facilitating networking between followers, or between the organisation and existing or potential new followers - helping to develop online and offline relationships through dialogue (Weber, 2009). The content of the tweets identifies causes the organisation supports such as national commemorative days i.e. 'Dignity Action Day' (Healthcare company), as well as new products/services. Individuals, such as employees are also introduced through the inclusion of links to interviews with senior members of staff, including the chief executive and the HR director (Healthcare company).

The focus of the next section is on the second stage of data collection, through the case-studies, semi-structured interviews with organisational storytellers. This section will also discuss sampling, access and participation. In contrast the objectives for the second stage of data collection, particularly focus on making sense of the online narrative through the process of narrative fusion (where elements of personal identity, and organisational identity are combined through narrative by the organisational storyteller and represented in their tweets as 'narrative public voice'). Therefore this stage requires a different methodological approach because it is the richness of the organisational storyteller's experience through narration of the process of writing the tweets that is required here.

12 Phase two – Antenarrative processes: Narrator – (Narrativist) paradigm

12.1 Qualitative semi-structured interviews

Following the identification of key themes from the tweet analysis, the second phase of the data collection was semi-structured interviews with the organisational storytellers from the five case-study organisations.

Given the relatively embryonic field of Twitter research from the perspective of those writing the tweets, particularly within the organisational context, this study's focus is a contemporary inquiry. Additionally, how individuals interpret and make sense of the online narrative in order to produce the tweets, may be described as a "complex social phenomena" (Yin, 2014:4) and so would benefit from further understanding, specifically in relation to online organisational narrative. Further details as to phase two of the data collection via semi-structured interviews is included below.

12.2 Characteristics of the participants

The semi-structured interviews were undertaken with 15 employees in total, who are involved in writing and responding to tweets on the organisation's Twitter site. The sample included eight men (53%) and seven women (47%). Within the five case-study organisations social media management is considered a specialism and often part of, though distinct from the marketing communications' function. Small marketing communications' teams are more common in B2B companies due to the emphasis on a more specific customer base and therefore more stringent targeted communication methods (Jensen, 2006).

12.3 Interviews: Sampling method, access and participation/purposiveness

Two to five employees were interviewed per organisation, the majority of these worked across management levels. The participants were interviewed once and their responses audio recorded and transcribed. A breakdown of interviewees per case-study organisation by gender and job title is included below in Figure 3.

The researcher's prior knowledge and experience of working within the social media function in the B2B sector enabled her to use established contacts to gain access into the

organisations initially. Introductions were then made from established contacts to those individuals working on updating the organisational Twitter site. A purposive sampling approach was therefore adopted as the individuals being interviewed were the organisational storytellers, or they have input into the Twitter creation process. It is only in this instance that these individuals can provide data on the sensemaking processes involved in identity construction – self and organisational – as well as provide insight into the influence of technological channels on creation of a narrative public voice. The approach to sampling undertaken follows Glaser's (1978:37) description of selective sampling as the:

> "*Calculated decision to sample a specific locale according to a preconceived but reasonable initial set of dimensions (such as time, space, identity or power) which are worked out in advance...*"

This informed approach to sampling allowed a structured interviewing timetable, where individuals were seen on the same day, so as to maximise the momentum immediately preceding the collection of tweet data. Each interviewee was asked to recommend others in the organisation that s/he felt the researcher should talk to. Brown & Thompson (2013) in their research on sensemaking, also draws attention to the necessity of talking to several members of the same team in order to understand the discrepancies and how these are translated into individual and team action. Different interpretations and responses to the same or similar processes is certainly a key element of the sensemaking process in this study, which is why at least two people per team are being included within the interviews. It is therefore the role of the researcher to draw these interpretations together from the tweets and also from the organisational storyteller's qualitative interview narrative.

As summarised by Dolores and Tongco, (2007) this is a non-randomised method where the researcher understands some of the information to be investigated, and then they choose the appropriate contacts based on their expertise in this field of study (Lewis & Sheppard 2006). Criticisms of sampling strategies in qualitative research are common, as Coyne (1997) summarises due to poorly explained assumptions, blurring of methods and theoretical standpoints. These criticisms do not apply directly to this study as the identification of roles and key contacts is based on experience and informed research of the selected case study organisations. An interpretivist qualitative approach is the most suitable approach as the depth of day-to-day experiences from those updating and monitoring Twitter sites within organisations is sought.

The main reason for conducting the interviews is to increase the depth of meaning from the tweet content, specifically around the process of identity construction and voice. In particular this part of empirical data collection will help to answer the second and third objectives for this study. These focus on making sense of the online narrative from the organisational storyteller's perspective and to understand the implications for them in producing the tweets, and how this positions the customer. Based on what they say about their day-to-day jobs, the researcher through the methodology is asking the organisational storyteller to make sense of how they interpret and communicate the organisational story. How the researcher interprets the 'story spaces' (or antenarratives) and how they in turn shape organisational and individual (organisational storyteller) identity is the next stage of the process. Followed

by how this blended narrative identity (narrative fusion, a mix of organisational and personal identity through narrative) is communicated through voice.

12.4 Exploring themes from the tweets written by the interviewees to inform the interviews

The aim of this stage of the research was therefore to explore the meaning behind the concepts inferred from the tweet data. To this effect the questions and tweet data are included as prompts – see also Figure eight for examples, to elicit further detail, giving interviewees the chance to explicate and expand upon the sensemaking of their specific phenomena (Saunders et al, 2009). It is though worth noting for the purposes of this study that preparation has been central to ensure the validity of the data gathered. This is particularly relevant here as individual organisational tweets, images and video clip 'stills' were downloaded and printed out (prepared) on pages to be shown to the interviewee, in order to elicit detail about the process of managing tweet content. In addition, a list of key headings, questions and tweet data has been collated in advance to suit the specific case-study organisation each time. The fact that there are only small teams working closely together on social media functions, including Twitter, within organisations meant that shared meanings and experiences were common among interviewees, which also helped to strengthen the value and reality of their accounts. Reality or lived experience in this sense, relates to the constructionist tradition of an agreement of opinions, as to what can be understood as truth (Guba and Lincoln, 1989) which is elaborated upon in the next section which focuses specifically on the emerging themes from the data.

13 Emergent thematic identification and reasoning – interpretive interview data Analysis

The analysis of the semi-structured interview data was also undertaken through an abductive approach. The process included the practice of reviewing the existing literature in conjunction with the empirical data, so as to explicate the importance of identity, narrative and voice in this research and to expand current thinking in this field (Van Maanen et al, 2007). In particular the aim of the analysis is to understand the narrative journey of the organisational storyteller through a process model that explores the stages and influences within tweet creation (Langley et al, 2013) from their experiences. The emphasis within this paper (Langley et al, 2013) is an examination of phenomena over duration and specifically how it evolves, alters and develops, there isn't quite the same emphasis on turning points and how you can identify these from the data within this study. Rather here the emphasis is on process research focuses on sequential developments of actions as the means of more easily making sense of them. For example the process of writing the tweets is an emerging phenomena that is examined here from the basis of the tweets themselves, as well as what the authors (organisational storytellers) say about the process of writing them. Some of the interview data contributes to this notion of organisational stories re-emerging as antenarratives through

the tweets, which provides specific insights into 'how the narrator makes sense of the past, work/organisational experiences.

13.1 Coding the interview data

Having identified the core interview questions through the initial tweet analysis, these formed the basis of the qualitative semi-structured interviews with the organisational storytellers (narrators). The emphasis of the abductive interpretive approach, requires "a constant movement back and forth between theory and empirical data" (Wodak, 2004:188, as cited in Mantere and Ketokivi, 2013). In this way, awareness and interpretation is seen as an 'ongoing conversation' between the existing theories and the researcher's previous experience. The process of referring back to the theory (relating specifically to identity narrative and voice) and looking ahead at the empirical data, should therefore inform a better understanding of the phenomena being studied, as well as providing new insights/knowledge (Mantere and Ketokivi, 2013; Martela, 2011). The interviews were therefore pertinent to understand how the organizational storytellers (narrators) make sense of their work/organisational experiences, in a manner described below by Butler (2005):

> "*. . . it is possible to explore how narrators make themselves subject to their new media narratives (Butler, 2005), allowing insight into the ways in which the possibilities and expectations of new media existence are internalised and with what implications for self and social relations.*"

It is this understanding of internalisation that is important here, because as previously discussed, just observing the writing of tweets will not provide sufficient depth to understand the implications of the organisational storyteller's identity construction. Also, part of the interview process was to test out a new concept of voice over the internet, initially conceptualised as 'narrative public voice'. This is because there are various different competing demands on the organisational storyteller when they are communicating via Twitter, one element is the challenge of maintaining the consistency of one organisational identity narrative. This is what current marketing literature (Nandan, 2005) suggests the organisational storyteller should be striving for to secure brand consistency.

In order to illustrate the manual coding of interview data, following initial thematic analysis a short extract from an interview with the marketing communications manager at Grant Thornton (business support) is included below as Figure 5.

As previously discussed the researcher is still in the process of analysing the data, so the initial findings and discussion below are focused on one case study only at this stage.

14 Emergent findings and initial discussion

Initial analysis focuses on an understanding of how many of the tweets fit into each of the categories: *Mobility, Accessibility, Interactivity, Visuality and Customisation*, as outlined by Elliott and Robinson (2013). The data from the 15 semi-structured interviews with marketing

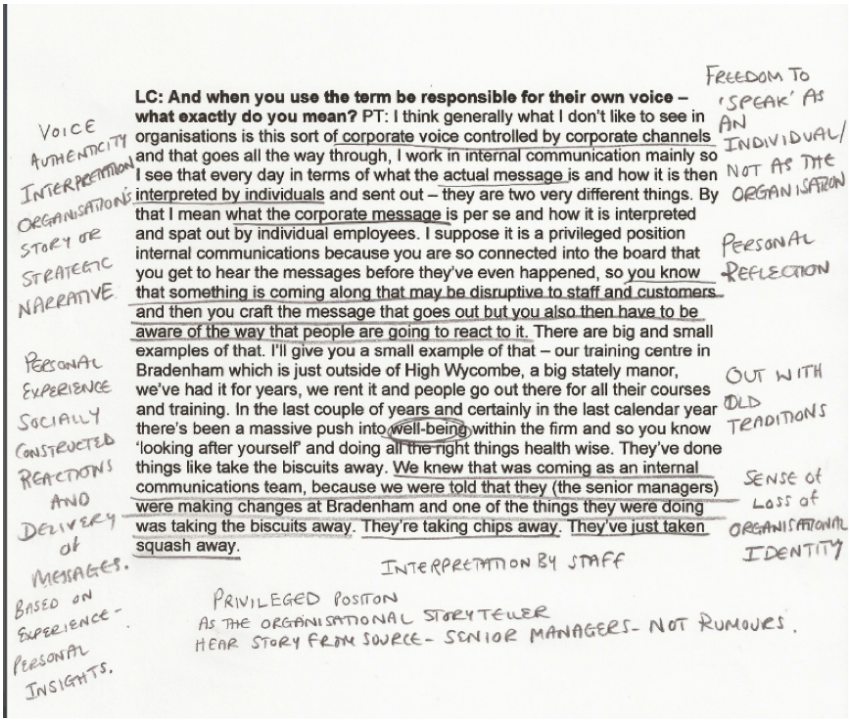


Fig. 5: The manual coding of interview data, following initial thematic analysis.

professionals has also been grouped into these categories and is incorporated to explain the intention of the organisation and the individual storytellers. As the author is at the early stages of data analysis, an initial review of the findings from one case study (manufacturing) is provided below.

Over a three-month period 54 tweets were collected from Forsdyke Industrial's Twitter site, a UK based manufacturer of air-technology and ventilation systems for buildings. The majority of the tweets, 15 (28%) were focussed on facilitating mobility and accessibility 15 (28%). An example of a tweet that specifically encourages the mobility of the customer from the main Twitter site to other linked articles were predominantly PR initiated stories relating to new product launches, for example:

```
Tweet: ‘‘Check out the front page of @MBS_magazine #ahu #sheffieldhallam
#recooler  #afreshapproach edition.pagesuite-professional.co.uk//launch.aspx?e...’’
```

While tweets that facilitate accessibility were principally those that advertised new vacancies and therefore were inviting the customers to become part of the internal team. In this sense access is being facilitated to the organisation itself, where customers have the opportunity to become part of its core identity, as staff members. Customers are informed within the tweet, which positions are currently vacant, and provided with a link to the job specification and the application form, also as detailed below:

```
Tweet: ‘‘Forsdyke Industrial UK currently has a vacancy for a Presentation
```

`Specialist & e-Learning Coordinator #jobs forsdyke.co.uk/newsarchive24/...''`

By narrating the story of the organisation positively, i.e. 'this is a good place to work' the marketing professional is also narrating their self, to an extent, 'I am proud to work here' but in a precarious context where reception is uncertain, as it is dialogic, at some level, and also potentially dependent upon the reception of an audience for its validation. Accessibility with the customer is sought in this context through validation where customers viewing the tweets may also want to work for the organisation and therefore will apply for a position within it. The organisational storytellers are also prepared to adapt their 'voice' in order to allow the organisation and themselves to seem more accessible. As Rebecca Williams, Marketing Assistant, highlights:

> "*When I write I aim to be the voice of 'the guy next door' because I think that's what our customers need to hear and what they recognise and I imagine you get better accessibility when you do that and because of the types of things that we use Twitter for, you know like product information, recruitment – so I don't think that if it was really the 'voice' of a marketing person it would have much legibility with people.*"

In this sense the presentation of self is performed to create the specific voice of an individual the customer will identify with. The guy next door is therefore considered more accessible than a marketing person in the conversational context of a tweet. This perhaps explains why there is less emphasis on tweeting key corporate messages. Only five out of the 54 tweets (9%) were categorised as 'Customisation and communication of strategic messages' which was the lowest category. Visuals and rhetorical questions are more frequently utilised to increase accessibility and to an extent interactivity, though this is very rarely reciprocated by customers as minimal 'retweets' and 'favourites' are noted. The visuals are principally focused on new products, visits from local dignitaries, such as MPs to the factory, 'Thanks for the retweets' messages and coming soon teaser campaigns. Sometimes these images are also accompanied by rhetorical questions such as:"My best RTs this week came from: @AFE_Dist_UK #thanksall Who were yours? Sumall.com/thankyou. Arguably though engagement with customers is not the principal aim of using Twitter in the B2B sector, as raising awareness may be a more achievable outcome, as John Griggs, Regional Operational Marketing Manager, explains:

> "*I see it (Twitter) as a way of raising awareness of the brand. The problem with B2B social media or any sort of digital marketing is really just about generating leads, you'll never actually get anyone to say: 'You're great – can we buy your vans, or stuff like that.' It's just to make sure people are aware as to who you are.*"

Perhaps one of the main reasons also that awareness of the brand seems a more likely aim than engagement, is because the organisation does not know who their followers are as John outlines:

> "*But we do try to convey the same sort of message online, in theory it is the same sort of people, at least we think it is, as we don't know exactly who our followers are.*

We have an idea of the groups from the retweets. So, I will go round looking for more of the right sort of people and follow them and hope they return the favour – they don't always, but I'll keep trying."

In this instance reiterating the identity of the brand consistently via Twitter is therefore more achievable. It also contributes to the notion of corporate identity as focussed on those aspects of an organisation that make it distinct, and therefore identifiable to customers and employees alike (Balmer, 1998, 2001, 2008, Christensen and Askegaard, 2001).

Although far more work is required to gain a more detailed understanding of the subject position of the organisational storyteller and the customer within B2B Twitter accounts, these very early findings suggest there is an emphasis on raising awareness of the brand, rather than engagement. A brief analysis of a manufacturing case study based in the east of England highlights an emphasis on moving customers away from the Twitter site through links is preferred to directly encouraging customer interaction. This is partly due to the nature of the identity of the organisation and also the content of the tweets, which is primarily factual and information-based rather than provocative and/or stimulating of discussions.

15 Contribution of this study, a new theoretical construct

One of the contributions to research this study offers is to develop a new theoretical construct focussing on the concept of a single unitary identity expressed through voice within the online organisational environment. Voice is being investigated as the principle way for organisational storytellers to combine their self-narrative and the official organisational story. This enables the organisational storyteller to effectively 'tell' the official story, in a way that encourages them to retain elements of their self-identity, while also effectively communicating the organisational narrative in a way that can be understood by customers. The 'narrative public voice' model has been re-examined during the fieldwork process for this study and the findings are in the process of being analysed and the model finalised.

Table 3: Breakdown of Interviewees by organisation, job title and gender.

Case Study	Organisation	Job Titles	Ages (in order of job titles)	Male/Female	Education Level
1	Telecommunications	Head of Customer, Insight and Futures, Head of online adoption and social media, Customer experience consultant, Social media lead and Marketing executive	50, 38, 35, 47, 27	4 Women, 1 Man	Doctorate, Degree, Degree, A-levels, Degree
2	Financial Services	Senior manager, national communications, Social media manager	38, 46	2 Men	Degree, Degree
3	Healthcare	Head of Digital, Communications manager, Group Communications Manager, PR manager and	34, 49, 31	2 Men, 1 Woman	Degree, A-levels, Degree
4	Business Support	Social media manager and Customer services manager	48, 45	1 Man, 1 Woman	Degree, A-levels
5	Manufacturing	Regional operational marketing manager, PR manager and Marketing assistant	49, 42, 28	2 Men & 1 Woman	Apprenticeship, post-school, Degree, Degree

Bibliography

[1] Aaker, D.A. (1986) Managing assets and skills: the key to sustainable competitive advantage. *California Management Review* 31 (2), pp: 91-106.

[2] Addullah, Z., Nordin, S.M., and Azis, Y.A. (2013) Building a unique online corporate identity. *Marketing Intelligence and Planning* 31 (5) pp: 451-471.

[3] Alvesson, M., Ashcraft, K. L., & Thomas, R. (2008). Identity matters: Reflections on the construction of identity scholarship in organization studies. *Organization*, 15, pp: 5-28.

[4] Aula, P (2010) Social media, reputation risk and ambient publicity management. *Strategy and Leadership* 38 (6) pp: 43-49.

[5] Balmer, J.M.T., and Wilson, A. (1998) Corporate identity: There is more to it than meets the eye. *International Studies of Management and Organisation* 28 (3) pp: 12-31.

[6] Balmer, J.M.T. (1998) Corporate identity and the advent of corporate marketing. *Journal of Marketing Management*, 14 (8), pp: 366-82.

[7] Balmer, J.M.T. (2001) Corporate identity, corporate branding and corporate marketing: Seeing through the fog. *European Journal of Marketing* 35 (3/4), pp: 248-291.

[8] Balmer, J.M.T. (2008) Identity based views of the corporation: Insights from corporate identity, organisational identity, social identity, visual identity, corporate brand identity and corporate image. *European Journal of Marketing* 42 (9/10), pp: 879 – 906.

[9] Barthes, R (1966/1977) "Introduction to the structural analysis of narratives" in Roland Barthes, *Image-Music-Text* (trans. Stephen Heath). Glasgow: Collins pp:79-124.

[10] Belenky, M. F., (1997) *Women's Ways of Knowing: The development of self, voice and mind*, Philadelphia: Basic Books.

[11] Boje, D. M. (2011). *Shaping the Future of Storytelling in Organizations: An Antenarrative Handbook*. London: Routledge.

[12] Boje, D. M., (1995) Stories of the storytelling organisation: A postmodern analysis of Disney as "Tamara-Land." *Academy of Management* 38 (4) pp: 997-1,035.

[13] Boje, D (1991) The story-telling organisation: a study of story performance in an office-supply firm. *Administrative Science Quarterly*, 36: 106-26.

[14] Braidotti, R (2013) *The Posthuman*. Cambridge: Polity Press.

[15] Brown, A.D., & Thompson, E.R. (2013) A narrative approach to strategy-as-practice. *Business History* 55 (7) pp:1143-1167.

[16] Brown, A.D., Stacey, P., and Nandickumar, J (2013) Making sense of sensemaking narratives. *Human Relations* 61 (8) pp: 1035 – 1062.

[17] Bryman., A & Bell, E (2011) *Business Research Methods*. Third Edition. Oxford: OUP.

[18] Budd, J. W., Gollan, P.J., and Wilkinson, A. (2010) New approaches to employee voice and participation in organisations. *Human Relations* 63 (3) pp: 303-310.

[19] Bullingham, L, and Vasconcelos, A.C (2013) The presentation of self in the online world: Goffman and the study of online identities. *Journal of Information Science* 39 (1) pp: 101-112.

[20] Butler, J (2005) *Giving an account of oneself*. New York: Fordham University Press.

[21] Butler, J. (1990; Anniversary edition 1999) *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

[22] Christiansen, L.T. and Askegaard, S. (2001). Corporate identity and corporate image revisited – a semiotic perspective. *European Journal of Marketing*, 35 (3/4), pp: 292-315.

[23] Cover, R (2012) Performing and undoing identity online: Social networking, identity theories and the incompatibility of online profiles and friendship regimes. *Convergence: The International Journal of Research into New Media Technologies* 18 (2) pp: 177-193.

[24] Coyne, I. T. (1997). Purposeful and theoretical sampling: merging or clear boundaries? *Journal Of Advanced Nursing*, pp: 623-630.

[25] Czarniawska, B. (2004). *Narratives in Social Science Research. Introducing Qualitative Methods*. London: Sage Publications.

[26] Davies, C. A. (2008) *Reflexive ethnography: A guide to researching selves and others*. Second edition. London: Routledge.

[27] Denzin, N.K., and Lincoln, Y.S (2005) *The Sage Handbook of Qualitative Research* (Third edition). Thousand Oaks, CA: Sage.

[28] Dijck van, J., (2013) You have one identity performing the self on Facebook and LinkedIn. *Media Culture and Society* 35 (2) pp:199-215.

[29] Dolores, Ma and Tongco, C (2007) Purposive sampling as a tool for informant selection. *Ethnobotany Research and Applications* 5 pp: 147-158.

[30] Eisenhardt, K.M., & Graebner, M.E (2007) Theory building from cases: opportunities and challenges. *Academy of Management Journal* 50 (1) pp: 25-32.

[31] Elliott, C. & Robinson, S. (2014). Towards an Understanding of Corporate Web Identity. In *The Routledge Companion to Visual Organization*. Bell, E., Warren, S. & Schroeder, J. London: Routledge. 273-288.

[32] Fetterman, D M (2010) *Ethnography Step-by-Step*. Third Edition. London: Sage Publications.

[33] Garton, L., C. Haythornthwaite., and B. Wellman., (1997) Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1). Retrieved September 19, 2000 from: `http://www.ascusc.org/jcmc/vol3/issue1/garton.htm`.

[34] Gergen, K.J (1991) *The saturated self, Dilemmas of identity in contemporary life*. New York: Basic Books.

[35] Glaser, B., 1978. *Theoretical Sensitivity*. Sociology Press, Mill CA: Valley.

[36] Goffman, E. (1959) *The Presentation of Self in Everyday Life*. London: The Penguin Press.

[37] Guba, E.G. & Lincoln, Y.S. (1989). *Fourth generation evaluation*. California: Sage Publications.

[38] Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice* (2nd ed.). London: Routledge.

[39] Harré., Rom (1998). *The singular self. An Introduction to the Psychology of Personhood*. London: Sage Publications.

[40] Hayles, K. H (1999) *How we became post human*. London: The University of Chicage Press.

[41] Hopfl, Heather (2002). Playing the part: Reflections on aspects of mere performance in the customer–client relationship. *Journal of Management Studies* 39 (2): 255–67.

[42] Jacobi, E., Freund, J., Araujo, L. (2015) Is there a gap in the market or a market in the gap?: How advertising planning performs markets. *Journal of Marketing Management* 31, (1-2), pp. 37-61.

[43] Java et. al (2007). Why We Twitter: Understanding Microblogging Usage and Communities. *Proceedings of the Joint 9th WEBKDD.*

[44] Jensen, M.B., (2006) Characteristics of B2B adoption and planning of online marketing communications. *Journal of Targeting, Measurement and Analysis for Marketing* 14 (4) pp: 357-368.

[45] Karreman, D and Alvesson, M (2001) Making newsmakers: Conversational identity at work. *Organisation Studies* pp: 59-89.

[46] Kozinets, R.V. (2010) *Netnography: Doing Ethnographic Research Online.* London: Sage Publications Ltd.

[47] Krishnamurthy, Gill and Arlitt (2008). A Few Chirps About Twitter. *Proceedings of the First workshop on Online social networks.*

[48] Langley, A (2013) Process studies of change in organisation and management: Unveiling temporality, activity, and flow. *Academy of Management Journal* 56 (1) pp: 1-13.

[49] Lawler, S (2008) *Identity: Sociological Perspectives.* Cambridge: Polity Press.

[50] Rosenberg, S. (1997). Multiplicity of selves. In R. D. Ashmore & L. J. Jussim (Eds.), *Self and identity: Fundamental issues* (pp. 23 – 45). New York: Oxford University Press.

[51] Lewis JL, Sheppard SRJ (2006). Culture and communication: can landscape visualization improve forest management consultation with indigenous communities? *Landsc. Urban Plan.* 77:291-313.

[52] Mantere, S and Ketokivi, M (2013) Reasoning in Organisations Science. *Academy of Management Review* 38 (1) pp: 70-89.

[53] Martela, F (2011) Abductive Mode of Inquiry – A Pragmatic Alternative For Conducting Organisational Research. Full paper submitted to *Subtheme 18, Pragmatism, Organising and Learning*, of the 27th EGOS Colloquium, Gothenberg, Sweden, 7-9 July, 2011. Available at: `http://www.frankmartela.fi/bibliography/DissertationAalto/Intersubjectivity_Abductive110601_Egos.pdf` Accessed on: 31 December, 2014.

[54] Markham, A. (2005). The politics, ethics, and methods of representation in online ethnography. In Denzin, N. & Lincoln, Y. (Eds.). *Handbook of Qualitative Research,* 3rd Edition (pp. 793-820). Thousand Oaks, CA: Sage.

[55] Marwick, A. (2013). "Ethnographic and Qualitative Research on Twitter." In Weller, 1 K., Bruns, A., Puschmann, C., Burgess, J. and Mahrt, M. (eds), *Twitter and Society.* New York: 2 Peter Lang, 109-122.

[56] McAdams, D. P, McLean, K. C (2013) Narrative Identity Current Directions in *Psychological Science* 22 (3) pp: 233-238.

[57] McPherson, T. (2000) 'I'll Take My Stand in Dixie-Net: White Guys, the South and Cyberspace', in B.E. Kolko, L. Nakamura and G.B. Rodman (eds) *Race in Cyberspace*, pp. 117–32. New York: Routledge.

[58] Murray, J.K (1995) Buddishm and Early Narrative Illustration in China. *Archives of Asian Art*, 48 17-31. JSTOR. IIt Bombay Lib., Mumbai, Maharashtra.

[59] Nandan, S. (2005). An exploration of the brand identity-brand image linkage: A communications perspective. *Brand Management*, 12 (4), 264–278.

[60] Purcell J and Hall M (2012). Voice and Participation in the modern workplace: Challenges and Prospects, *Acas Future of Workplace Relations series* Available at: `http://www.acas.org.uk/media/pdf/g/7/Voice_and_Participation_in_the_Modern_Workplace_challenges_and_prospects.pdf` Accessed on: 19 February, 2015.

[61] Rosile, G.A et al (2013) Storytelling Diamond: An antenarrative integration of the six facets of storytelling in organization research design. *Organisational Research Methods* 16 (4) pp: 557-580.

[62] Saunders, M., Lewis, P and Thornhill, A (2009) *Research methods for business students.* Fifth Edition. Harlow: Pearson Education Ltd.

[63] Sawhney, M, Verona, G and Prandelli, E (2005) Collaborating to create: The internet as a platform for customer engagement in product innovation. *Journal of Interactive Marketing* 19 (4) pp:4-17.

[64] Sillince J.A.A., & Brown A.D. (2009) Multiple organisational identities and legitimacy: The rhetoric of police websites. *Human Relations* 62(12) pp:1829-1856.

[65] Tajfel, H and Turner, J. C (1979) An integrative theory of intergroup conflict. In W.G Austin and S. Worschel (Eds), *The Social Psychology of Intergroup Relations* (2nd Ed, pp: 7-24) Chicago: Nelson-Hall.

[66] Turkle, S (1995) *Life on the screen.* London: Simon and Schuster Paperbacks.

[67] Tyler, M. (2011) 'Tainted Love: From Dirty Work to Abject Labour in Soho's Sex Shops', *Human Relations*, 61 (11).

[68] Valle, V.M and Torres, R. D (2000) *Latino Metropolis* Minneapolis: University of Minnesota Press.

[69] Van Maanen, J., Sørenson, J., and Mitchell T.R. (2007) The interplay between theory and method (Editors introduction to Special Topic Forum). *Academy of Management Review*, 32, (4), pp: 301-314.

[70] Vickers, M.H. (2005) Illness, work and organisation: Postmodern perspectives, antenarratives and chaos narratives for the reinstatement of voice. *Tamara: Journal of Critical Postmodern Organisation Science* 3 (2) pp: 74-88.

[71] Watson, T. J. (2008) Managing identity: Identity work, personal predicaments and structural circumstances. *Organisation* 15 (1) pp: 121-143.

[72] Weber, L (2009) *Marketing to the social web.* Second edition. New Jersey: John Wiley & Sons Inc.

[73] Webster's Dictionary (1966) *Webster's Third International Dictionary.* Massachusetts: Merriam-Webster.

[74] Weick, K. E (1995) *Sensemaking in Organisations.* London: Sage Publications.

[75] Williams, M (2000) Interpretivism and Generalisation. *Sociology*, 34 (2) pp: 209-224.

[76] Wieland, S. M. B. (2010) Ideal selves as resources for the situated practice of identity. *Management Communication Quarterly* 24 (4) pp: 503-528.

[77] Yin, R.K (2014) *Case Study Research: Design and Methods*. London: Sage Publications Ltd.

Twitter data for urban policy making: an analysis on four European cities

Marta Severo[1], Timothée Giraud[2], and Hugues Pecout[3]

[1] Laboratoire Geriico GIS-CIST
Université de Lille 3
Villeneuve-d'Ascq, France
marta.severo@univ-lille3.fr
[2] CNRS, UMS Riate
timothee.giraud@ums-riate.fr
[3] CNRS, GIS-CIST
hugues.pecout@gis-cist.fr

Abstract. This paper presents the results of a project (called "Big data") carried out in the context of the European programme ESPON. In this project, we test the interest of using Web 2.0 data for studying issues related to territorial development and cohesion. We focus on the possible uses of large sets of Twitter data for studying European cities. These data are expected to improve the study of the impact of the brand of a city at international and at local level. The objective of the project is to define a methodology and not to provide final evidence. For this reason, we test three different types of samples and methods in order to evaluate advantages and drawbacks of each technique: (first method) tweets mentioning the city; (second method) tweets geo-tagged in the city; (third method) tweets of a qualitative sample of influencers from a city.

Keywords: Twitter, city, brand, visualisation, communication, public policy.

1 Introduction

This paper presents the results of a project (called "Big data") carried out in the context of the European programme ESPON[4]. In this project, we test the interest of using Web 2.0 data for studying issues related to territorial development and cohesion. We focus on the possible uses of large sets of Twitter data for studying European cities. Twitter is an online social networking and micro-blogging service that enables users to send and read short messages called "tweets". Today this platform has more than 200 millions active users and more than 500 millions tweets sents per day. There are three features that make tweets very interesting for our analysis: (i) tweets are short (ii) they are generally public and can be easily collected through the API (iii) they are or can be geo-tagged. These data are expected to improve the study of the impact of the brand of a city at a local and international level.

[4] http://www.espon.eu/main/Menu_Calls/Menu_Procurements/Menu_PreviousProcurements/Feasibility_Study_Analytical_Tools_Big-Data.html

The objective of the project is **to define a methodology and not to provide final evidence**. For this reason, we test three different types of samples and methods in order to evaluate advantages and backwards of each technique: (first method) tweets mentioning the city; (second method) tweets geo-tagged in the city; (third method) tweets of a qualitative sample of city's influencers. We analyse tweets related to four European cities (Marseille, Brussels, Edinburgh and Bologna) for a period of four weeks. In the first part, we will present the methodology that we have designed for this case study. In the second part we will describe the results of the analysis and in the third part, we will summarize our conclusions based on the project.

2 Method

We analyse tweets concerning four cities (Marseille, Brussels, Edinburgh and Bologna). Data available concern the period from 21st May to 6th July[5]. We focus especially on a period of four weeks (2-29 June 2014). We propose to select these metropolitan areas because they: guarantee a sufficient coverage of the ESPON area; constitute very good examples of metropolitan projects; have similar size (between 1 and 3 millions habitants); guarantee a good and feasible linguistic and cultural variety.

As said, the objective of this study is to define a methodology and not to provide final evidence. For this reason, we decide to test three different methods of using Twitter in urban studies and we evaluate the advantages and drawbacks of each method. In the first method, by analysing tweets including the name of the city, we expect to identify the items that constitute the image of a city. In the second method, by analysing tweets geotagged from the four cities, we will give insights on the intra-urban activities of the city. Then, in the third method, by identifying a sample of influencers on Twitter for each city and by analysing their tweets, we expect to observe in real-time the themes related to the metropolitan area.

Yet, before presenting the analysis's results, there are two main drawbacks related to Twitter data that have to be taken into account: (1) the variety of uses of tweets and (2) the small amount of geo-tagged tweets.

2.1 The variety of uses of Twitter

Twitter is a tool with several different uses. On the one hand, people write tweets for expressing emotions, for communicating or organizing their activities, for spreading their ideas. On the other hand, Twitter can be used as a promotional tool for diffusing events, products or buzz. Moreover, several websites or social network accounts (Tumblr etc.) can use tweets for syndicating their content.

Another important issue to be considered is the impact of some events on the normal Twitter activity. If we consider tweets from the 2nd of June to the 29th of June 2014, some important peaks can be identified in time series of each city (except for Bologna). For example,

[5] Tweets have been collected thanks to the TCAT tool developed by the University of Amsterdam - Digital Method initiative (Rieder & Borra, 2014).

in Brussels, a peak on the 2nd of June is due to a demonstration related to the abdication of Juan Carlos (#ReferéndumYA, #IIIRepública, #ElReyAbdica), followed by the peak of the 3-5 June related to the G7 summit. In Edinburgh, there is a peak at the beginning of June related to the concert of the band "One Direction" (#NiallsNotes, #WWATour, #1DA). Different is the case of Marseille where there is a peak on the 21st of June that is not related to a few specific hashtags, but is a result of the intensification of the general activity on Twitter because of a cultural local event ("fête de la musique") that pushed people to tweet. Such phenomenon can create important biases, especially for aggregated analysis across time. Yet, it is very difficult to define a general rule for treating it.

As we will see in the following pages, it appears difficult to understand whether a tweet really concerns the city (even if the city is mentioned) and even more difficult to understand which content of tweets is related to the image of a city. No doubt, it is impossible to develop a standardized and automatic method for distinguishing pertinent and non-pertinent information.

2.2 The small amount of geo-tagged Tweets

Another important limit that has to be taken into account is the scarcity of geo-tagged tweets, usually no more than 1% of the total tweets (Gerlitz & Rieder, 2013). The geo-tagged tweets we will use in this study are those that contain an actual latitude/longitude coordinate pair, derived from the GPS sensor on a smartphone or through cell tower triangulation. In our study, tweets' geographical distribution is normalised with the distribution of a 1% random extraction from all tweets. Through this normalisation, we can focus on the over-quotation of cities instead of on the simple quotation.

Several other methods have been experienced to infer user's location on Twitter based on user profile descriptions, user profile languages or following-follower relationships (Davis et al, 2011; Crampton et al, 2013), yet these methods are still experimental and are really time-consuming if applied on larger corpora. For this reason, we decided to exclude them in this first experimental phase of the study.

3 Results

3.1 Frist method: tweets about the city

As a first method, we propose to analyse tweets mentioning the name of the four cities in all European languages. Through this technique, we expect to study the recognition of a city at the international level and to identify the items which are acknowledged as part of the brand of a city. On this sample we performed two types of analysis. First, we used geo-tagged tweets in order to analyse the geographical distribution of people mentioning the city. Second, we focus on the topics related to a city through different content analysis techniques, by combining quantitative and qualitative methods.

Spatial analysis: where are located the users mentioning the city? In a first phase, we focus on geo-tagged tweets to study where are located people speaking about a city. As said, we have to keep in mind that these tweets represent only a small percentage of all tweets mentioning the city: between 5% of Bologna and 2% of Edinburgh. We can observe two facts:

(1) Tweets that mention a city are more likely to be geo-tagged. They are from 2 to 5% in our sample whereas the general ratio is more around 1 or 1.5%.
(2) The bigger sample, the lower is the percentage of geo-tagged tweets (Table 1).

Table 1: Tweets and geo-tagged tweets mentioning the four cities from the 21st May and the 6th July.

City	**Total Tweets**	**Geo-tagged Tweets**	**Ration**
Bologna	233,899	12,089	5.17%
Brussels	739,150	20,183	2.73%
Edinburgh	1,064,735	21,583	2.03%
Marseille	430,824	18,344	4.26%

We analyse the geographical distribution of tweets mentioning a city. By comparing the city's quotation density around the world to the general density of tweets, we can easily see where the city is over-quoted in relation to the normal Twitter activity (Figure 1). Our hypothesis is that the geographical distribution of over-quotation could be used as an indicator that reflects the level of internationalisation of a city.

The geographical distribution of tweets mentioning the four European cities exhibits some regularities. Generally, we can observe that tweets that mention a city are mostly emitted from the city's country and even more from within the city and its suburban area. We can identify some exceptions:

- The over-quotations of Bologna in USA due to the fact that there is a city called Bologna also in USA and in a similar way in Spain where there is zone called Bolonia;
- The over-quotations of Brussels that are dispersed in several European countries. This is probably due to the European role of this city (in several case the term "Brussels" can become a synonym for "European Union");
- the over-quotations of Marseille in North Africa probably due to the geographical and cultural proximity with other Mediterranean cities.

So as a conclusion, the analysis of over-quotations' distribution can be useful as an initial exploratory technique for two goals. First, over-quotations' distribution appears quite a good indicator of the international recognition of a city. The different international role of Brussels compared to the other cities clearly emerges by using this indicator. Second, the identification of peaks of over-quotation outside of the city's country can help in detecting

trivial errors of homonymy that could be verified by content analysis such as in the case of Bologna in our sample.

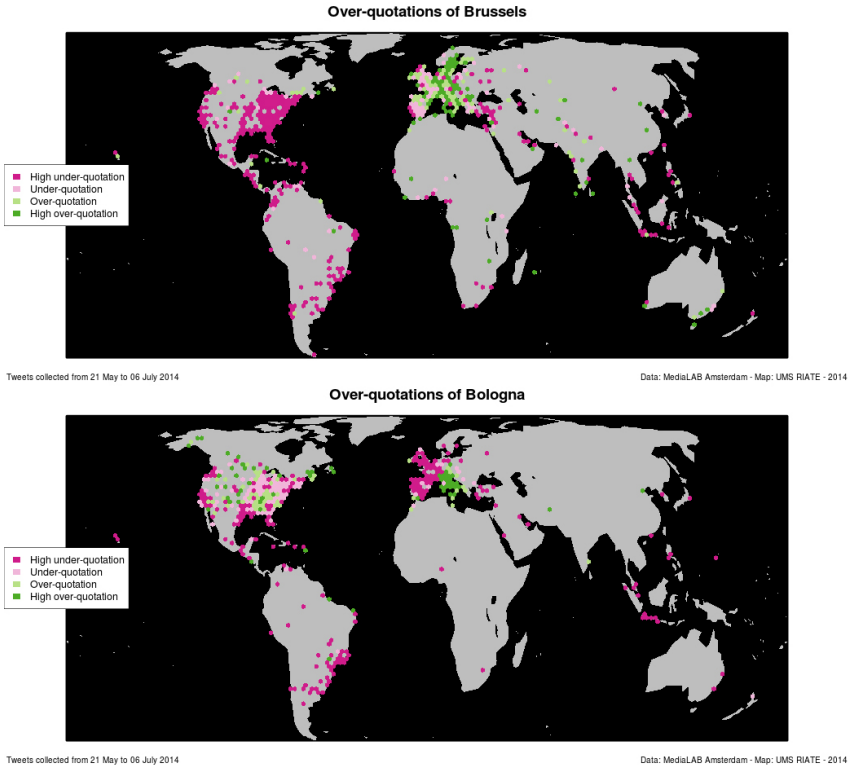


Fig. 1: Geographical distribution of over-quotations of the Bologna and Brussels at global level.

Content analysis: which are the topics associated to the city? In a second phase, we consider all tweets mentioning a city (not only the geo-tagged) in order to identify topics addressed by Twitter users in relation to the city. This technique is meant to identify the items of the brand of the city that are recognised all around the world. Considering the large amount of tweets for each city, only automatic techniques of analysis are feasible. Yet, it is important to underline that in content analysis of tweets, even when we are using an automatic technique, qualitative interpretative work is equally necessary. Often users employ abbreviations and special expressions that can be not so immediate for the researcher.

We start by considering the most frequent hashtags[6] for each city (Table 2). In order to evaluate the representativeness of hashtags, it is very important to consider how many distinct users mention them. In this project, we establish that when the ratio between the number of tweets mentioning the hashtag and the number of distinct users sending tweets

[6] Hashtag analysis is supposed to be more effective than simple word analysis for identifying the topic of a tweet. Indeed, tweets users employ hashtags to connect their tweets to a general discussion.

is higher than 10, we can exclude the hashtag because it can be considered as a form of advertisement and it does not reflect a shared discussion.

Once excluded non-representative hashtags, results are quite generic. If we take the first 20 hashtags in each city, we find several hashtags generated by special events and only some general tags related to the city. There are some common topics such as job and travel that emerge in every city, then for Bologna and Marseille, football clearly emerges as the main topic related to the city, for Brussels, the link with Europe is surely the main characteristic that we can identify. Conversely, for Edinburgh, there is no information except for the One Direction concert and the referendum in Scotland (#indyref) that can be detected with an aggregated analysis of tweets over a one-month period.

In order to avoid this problem of few large peaks across the aggregate data, we try to disaggregate by analysing the frequency of hashtags at the level of a week. Indeed, this technique is very useful to identify hashtags related to specific events through time. For example in the case of Bologna, we can detect the impact of the audition of the XFactor and the IT fair SMAU the first week, a convention of a political party the second week, the projection of the Lego Movie film the third week and the gay pride last week. Yet, this technique doesn't help us in our mission of finding items of the image of the city. There are very few contents that are constantly mentioned over the four weeks. We find again job, travel, football, Europe... These are still really generic.

Another possible way to carry out a thematic analysis consists in analysing the list of hashtags through a qualitative content analysis. As a first step, non-pertinent (for homonymy) and non-representative (ratio tweets/user) hashtags have to be excluded. Then, hashtags mentioned a sufficient number of times (in this project 50 times over the period of observation) can be classified in different categories composing the brand of a city. We propose to use 14 categories: crime, culture, economics, politics (national/local), international politics, social affairs, transport, urban planning and life, sport, tourism, food, fun, weather and places (other places mentioned in the same tweet with the city). This analysis can be partially automatised by building a dictionary that connects hashtags to categories.

The advantage of this technique is that it can provide a more detailed view of the city's topics that draw people attention on Twitter. Then, through a treemap, it is possible to easily visualise which hashtags are more used inside each category (Figure 3). This tool can be very useful for decision makers that can have a general point of view on topics related to the city and that can quickly identify hot topics that affect the image of the city at national and international level. If we consider the example of our four case studies, we can note that this technique can produce some interesting insight.

As an example, we will present the case of Bologna. In this corpus, we have a total of 299 hashtags used at least 50 times in 92,797 tweets. We exclude 32 hashtags because non-representative and 26 because non-pertinent.[7] Figure 2 shows the distribution of hashtags related to Bologna by category and Figure 3 shows the distribution of tweets by hashtag and hashtags' category through a treemap.

[7] We also excluded name of the city and the city's country.

Table 2: Most frequent hashtags in tweets mentioning a city (2-29 June). Red = possible errors; grey = non-pertinent; white = generic; green= events; yellow; = city related; blue = places.

BOLOGNA

hashtag	# of tweets	distinct users	ratio
bologna	31,720	10,056	3
bolonia	2,702	2,493	1
news	2,553	150	17
tarifa	2,351	2,261	1
andalucía	2,244	2,177	1
campogibraltar	2,231	2,162	1
cádiz	2,116	2,074	1
italy	1,256	817	2
xf8	976	554	2
lavoro	779	134	6
meteo	752	71	11
thelegomovie	743	741	1
ss16	740	14	53
smau	726	337	2
eritrea	575	133	4
italia	562	393	1
bolognafc	560	82	7
roma	553	253	2
calciomercato	528	123	4
seriea	493	155	3

BRUSSELS

hashtag	# of tweets	distinct users	ratio
Brussels	19,010	10,671	2
Bruxelles	12,285	6,903	2
Belgium	5,43	2,751	2
Bruselas	3,190	2,084	2
g7	3,039	2,271	1
Job	2,951	312	9
Jobs	2,791	344	8
Eu	2,621	1,927	1
catalanswanttovote	2,314	1,704	1
Ukraine	2,049	1,580	1
Inclusive	1,567	1,563	1
Innovative	1,554	1,550	1
News	1,508	385	4
Bel	1,262	1,181	1
Europe	1,186	813	1
Belgique	1,178	553	2
referendumya	1,086	930	1
It	1,065	49	22
Afp	1,058	877	1
Travel	952	577	2

Table 2: **(Continued)**. Most frequent hashtags in tweets mentioning a city (2-29 June). Red = possible errors; grey = non-pertinent; white = generic; green= events; yellow; = city related; blue = places.

EDINBURGH

hashtag	# of tweets	distinct users	ratio
edinburgh	82,200	28,735	3
new	19,955	5,364	4
wwa	11,339	11,141	1
wwatour	9,600	8,368	1
job	9,294	374	25
jobs	8,713	552	16
scotland	8,626	3,991	2
niallsnotes	7,870	1,752	4
1da	7,005	1,938	4
hq	6,126	2,435	3
nn	4,494	1,842	2
indyref	3,394	1,557	2
ns	2,764	1,443	2
gumball3000	2,516	1,161	2
onedirection	2,116	888	2
glasgow	2,047	822	2
niallspotato	1,867	567	3
london	1,774	814	2
travel	1,578	605	3
notasstyles	1,538	458	3

MARSEILLE

hashtag	# of tweets	distinct users	ratio
Marseille	33,756	13,049	3
Om	4,242	2,575	2
marseillelife	3,676	3	1,225
Teamom	3,663	2,327	2
France	3,507	1,179	3
Alg	1,735	1,612	1
cdm2014	1,529	1,285	1
olympique	1,498	1,145	1
Worldcup	1,397	1,339	1
Algrus	1,373	1,102	1
Football	1,104	878	1
Algerie	947	841	1
Paris	945	617	2
shake14	910	383	2
Lfc	807	769	1
Travel	796	548	1
Mufc	728	712	1
Marsella	676	448	2
Fcb	665	653	1
Mercato	648	536	1

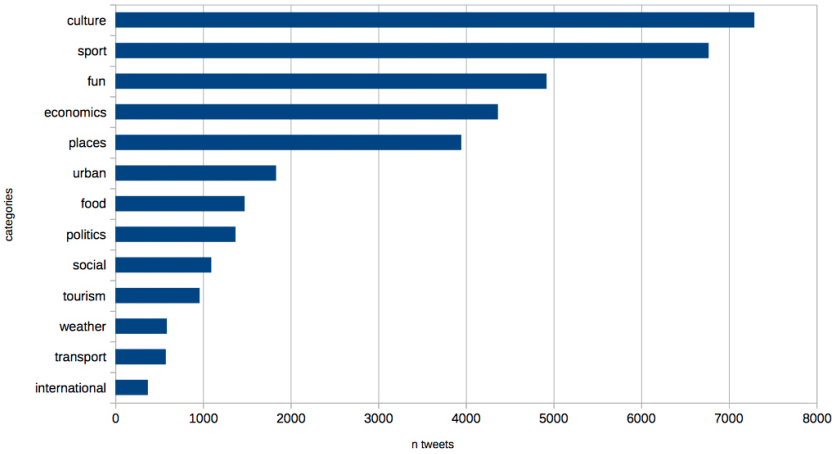


Fig. 2: Distribution of hashtag in tweets mentioning Bologna between the 2nd and the 29th June 2014.

Through this method, we can clearly confirm that culture and sport are the main topics related to Bologna. As regards "culture", most frequent hashtags are related to national and international events happening in the city (such as the audition of the musical TV program XF and the release of the Lego Movie). Yet, there are numerous local events (festival, concerts, cinema and art events) that are frequently mentioned. As regards "sport", football is a key topic, yet we can also see the effect of the World Cup.

If we focus on aspects related to spatial analysis and urban planning, several insights that could be useful to decision makers can be listed:

- Few places inside the city are mentioned: particular places and streets (#piazzamaggiore), cultural venues (#cassero), commercial spaces (#interporto), monuments and historic building (#sanluca). It is really interesting that the Portici (arcades), which are considered as one of the main symbols of Bologna, do not appear in the list of hashtags.
- Some urban questions emerge. For example, #Nopianocasa hashtag is related to a national social movement for the right of habitation against violent evictions.
- Items of the economic sector mentioned at international level. The SMAU, IT fair, is mentioned in several tweets. Then, we find the role of the university through several hashtags. Also the fashion sector is cited (in particular the local retailer yoox) and the motor sector with the hashtag #ducati.
- Very few content related to international politics. We find only some tweets related to a public mobilisation against the Europan meeting of the 11 July 2014 (#civediamolundici).
- People communicate about different types of transport, especially highways (#a1, #a14) and trains (#trenitalia), but also bicycles (#bici), flights (#aereo) and bus. What is interesting is the hashtag #viabilibo that allows to monitor the impact of road works.
- The hashtag #mybologna is used on Twitter to indicate what people like of Bologna. This hashtag can be a very precious source of information about the perception of the

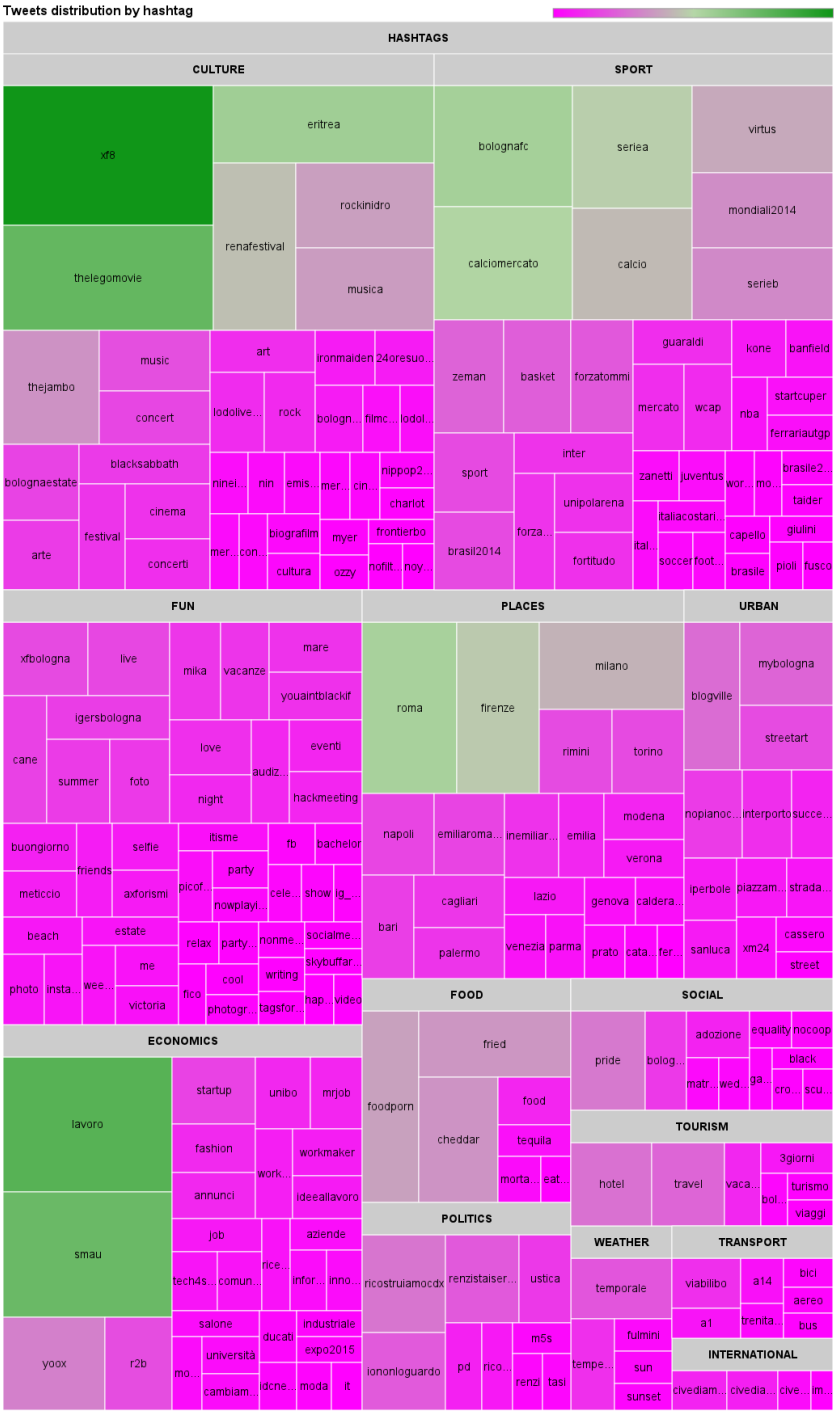


Fig. 3: Treemap of hashtags in tweets mentioning Bologna between the 2nd and the 29th June 2014.

city. 139 different users have used it 266 times over the four weeks. Co-hashtag analysis[8] is a useful quick technique to see the context of use of a hashtag (see Figure 4).

– Co-citation with other places: in numerous tweets (13,182), Bologna is cited with other places, notably cities such as Roma, Florence and Milan and several others. Yet, only Italian places are mentioned.

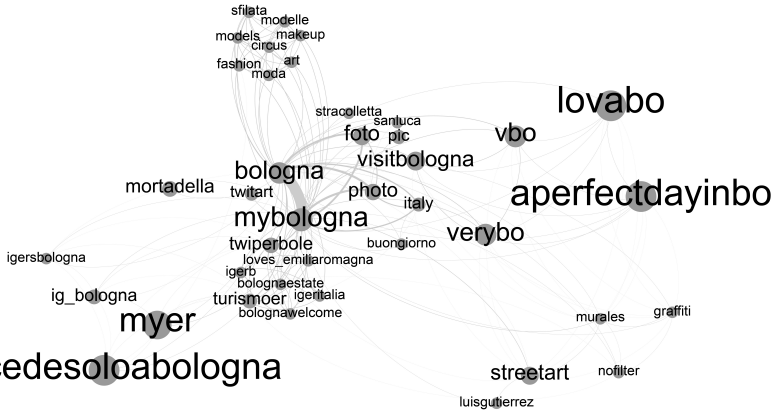


Fig. 4: Co-hashtag analysis of the hashtag #mybologna in tweets mentioning Bologna (2-29 June 2014). Size of nodes is proportional to word frequency

4 Tweets inside the city

Through the first method, we have two main findings: (1) geo-tagged tweets seem to contain more space-related and city-related content; (2) the majority of geo-tagged tweets are located inside and around the city. Given that, as a second method, we propose to focus exclusively on tweets geo-tagged n the metropolitan areas of the four cities.[9] Doing this, we expect to reduce the noise related to occasional events and find more spatial-related content. Moreover, the analysis of discourses of users located in city is expected to give insights of Twitter communication at local level. As in the previous case, first we focused on spatial analysis, then on content analysis.

Spatial analysis: where are located in the city the users tweeting? The sample is constituted of 105,188 tweets from 7,357 users for Bologna; 146,040 tweets from 8,941 users for Brussels; 193,547 tweets from 11,713 users for Edinburgh; 288,453 tweets from 25,730 users for Marseille. So, once again, it is necessary to underline that this is just a very small part of the global Twitter activity. As a first step, we can analyse the distribution of tweets in the city in order to see if there are some places where people tweet more than other. The

[8] The analysis of network of co-citation of hashtags in the same tweet.

[9] City areas have been selected according ESPON standards NUTS3 coordinates.

comparison with the population density helps us to identify urban places where Twitter activity is concentrated. Globally, we observe that the distribution of tweets is similar to the distribution of population inside all four cities, yet some places where Twitter activity is particularly intensive can be identified.

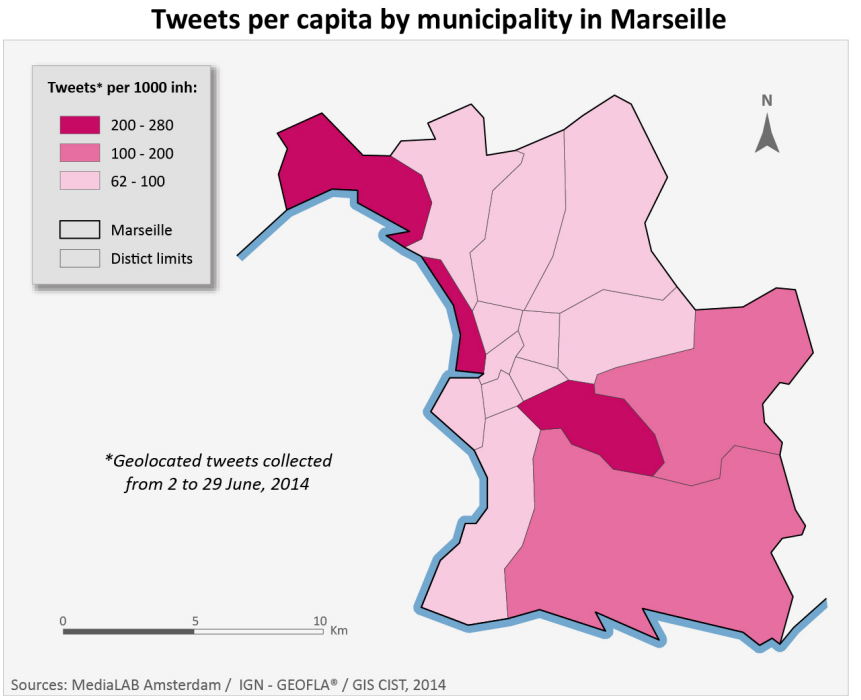


Fig. 5: Geographical distribution of tweets per capita in Marseille.

In general, we can see that tweets are spread out all around the city quite proportionally to the population density. Yet, in every city there are same places that emerge. For example, in Marseille (Figure 5) tweets are distributed in every part of the city, yet we can locate notably the old port, the stadium and the train station. At the borough level, we see a concentration (200-280 tweets per 1000 inh) in the 2nd borough were we can find a lot of touristic and cultural attractions as the new museum MUCEM, a new mall or the city hall. We can note that these sectors, like the 16th (North), concentrate a lot of young people. In Bologna (Figure fig:ch8-fig6) the situation is quite different. Tweets are concentrated especially in the downtown, in the area around the main square, Piazza Maggiore and along the main route that connects the city from North to South (via ugo Bassi – via Rizzoli). Moreover, we can identify some places where there is a strong Twitter activity, such as the airport, the fair, the hospitals and the stadium. Also university buildings clearly appear as tweeting places.

Content analysis: which are the topics associated to the city? As a second step, we try to identify main topics that are debated inside the city. Also with this corpus, most

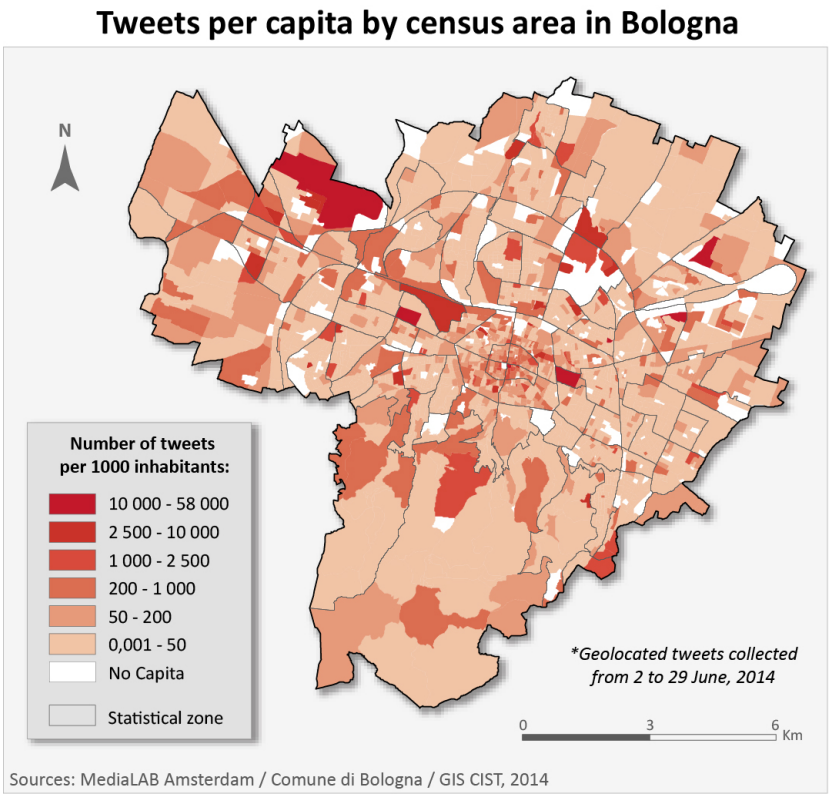


Fig. 6: Geographical distribution of tweets per capita in Bologna.

frequent hashtags are not useful to catch the diversity of the discussion. So, considering the good results of the first corpus, we use the same technique, quantitative analysis of hashtag, on the corpus of geo-tagged tweets inside the city. We focus again on the example of Bologna.

We have a total of 282 hashtags used at least 20 times[10] in 19,194 tweets. We excluded 69 hashtags because non-representative.[11] The results of this analysis are very unexpected. Even if tweets are geo-tagged, they often don't speak about space and even less about cities. In our corpus, 44% of tweets including hashtags talk about sport and notably about the world cup. The coincidence with this big event makes it very difficult to know if it is an accidental effect or it is usual that geo-tagged tweets don't speak of the city. Nevertheless, we can note that 25% of tweets are tagged as "fun": they are used principally to share emotions, photos or organise fun activities (parties, hang-out etc.). So, these tweets are surely related to space yet the communication often is just expressive and phatic. Conversely, there are very few tweets related to urban questions and tweets related to transport and international politics are completely absent. Figure 7 shows the repartition of tweets by category. We use the same categories as presented above.

[10] Considering that tweets geo-tagged in the city are less than tweets mentioning the city we take a lower level of hashtag frequency in order to have a comparable amount of hashtags.

[11] We also excluded name of the city and the city's country.

Based on this case study, we can make the hypothesis that geo-tagged tweets are not so useful for studying the city. Even if located in the city, they generally are very few (not representative of the general discussion) and don't concern urban life, but rather private life. The only possible exploitation that we see concerns tweets that mention photos. As an example, we verify the content of tweets including the following hashtags: #picoftheday, #photooftheday, #photo, #igers, #instamood, #igersbologna and #Instagood (the last four are related to the photo sharing application Instagram). In this way, we can build a corpus of 518 tweets including images of the city that can give us insight of the perception of some places. Figure 8 represents the word cloud generated by these tweets can provide an overview of the type of content of these messages.

4.1 City's influencers on Twitter

We have tested two methods so far. As a first method, we analysed tweets mentioning the name of the four cities in all European languages. This technique was useful to have an indicator of the recognition of a city at international level and to have a general point of view on topics related to the city. As a second method, we analysed tweets geo-tagged in the city. This method should help us to study more precisely the discussion related to the city, yet actually this corpus proved to be very general. People inside the city rarely speak about the city, most often their communication is emotional in nature.

In order to bypass this issue and to find more precise insights concerning urban life, we propose a third method. By identifying about 40 influencers on Twitter for each city (for example members of the city council, chamber of commerce, big local enterprises, local political leaders, local ONG leaders, blog stars...) and by analysing their tweets, we expect to observe in real-time the themes related to the city and the metropolitan area. This analysis will provide a representation of the public debate in the digital sphere, especially with a focus on the metropolitan issues that are emerging.

Selecting a specific sample of Twitter users is meant to avoid the noise related to sport and fun that we found in the previous corpora. As influencers, we privilege public people and institutions rather than media. For this test, influencers of each city have been chosen by an expert in urban planning that have lived in the city. If this experiment is successful, in a project at larger scale, influencers have to be chosen by a panel of decision makers and experts working in the city.

Considering that this sample of users is smaller than the two previous ones, it was possible to collect tweets on a longer period of one year from the 1 July 2013 to the 30 June 2014. For Bologna the corpus consists of 60,173 tweets sent by 44 influencers, for Brussels 32,592 tweets by 39 influencers, for Edinburgh 53,182 tweets by 42 influencers, for Marseille 46,208 tweets by 50 influencers. These tweets are not equally distributed along the period. By considering the time series of tweets in the four cities, two phenomena may be observed: (1) tweets are distributed according a weekly cycle: influencers usually tweet less the weekend. Moreover, in all cities there are some less active periods during August 2013 and Christmas holidays. This fact can be considered a good indicator of the link between the professional life of

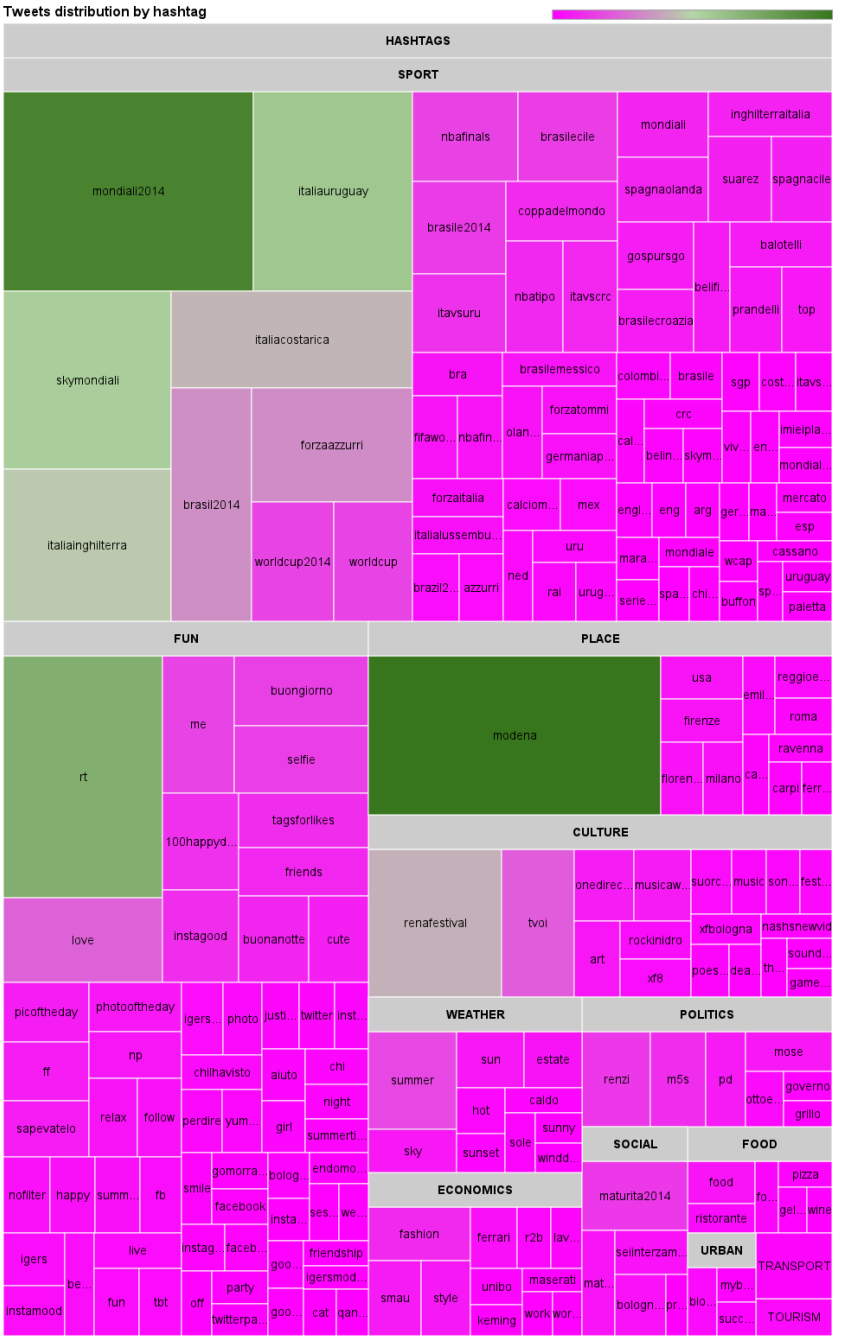


Fig. 7: Treemap of the hashtags used in tweets geo-tagged in Bologna between the 2nd and the 29th June 2014.

the influencer and the communication conveyed on Twitter. (2) Along the year, tweets are increasing.

We study this corpus principally through content analysis.[12] Once again, quantitative techniques are preferable considering the amount of data to study. In order to be effective and useful for decision makers, we propose a methodology organised in four phases:

1. General overview of the corpus. As a first step, it is possible to have a global idea of the contents of the discussions through the analysis of 50 most frequent hashtags.
2. Diachronic overview of the corpus. Then, visualising the most frequent hashtags per month through a stream graph will allow observing the evolution of the discussion. Decision makers can easily identify themes that occupy public debate and follow their development along time.
3. Focus on specific themes. Once identified an interesting theme, decision makers can obtain a summary of the discussion through three techniques:
   (a) Most retweeted messages is an quick technique to know the most popular contents that occupy the discussion;
   (b) Co-hashtag analysis allows to visualise how the theme is connected to others;
   (c) Word-frequency analysis (and the visualisation through word clouds) allows identifying the main elements that are mentioned in the debate concerning the selected theme.
4. Qualitative analysis of tweets. As a final step, decision markers may need to perform a close reading of the tweets themselves in order to have a complete view of the debate.

In the next paragraphs, we present some examples of this analysis workflow. Once again, we will focus on Bologna.

Even at a first look, the content of this corpus appears strongly different from the previous ones. By considering the most frequent hashtags along the year of observation (Annex 14), Sport remains a very important topic (first hashtags are #bolognafc the football team and #granarolo #virtus, the basket team), yet "fun" is practically inexistent. "Unibo", the hashtag of the University of Bologna, is largely used and then there are several keywords related to politics (#ilgiornodopononsuccedemaiuncazzo related to political scandals, the main political parties #pd #ms5 and a few mentions for the right party #pdl) and social life, notably social movements (#19o for the strike of 19 October 2013, #occupymensa for the student strike, #sollevazione and #assedio generally related to social demonstrations).

If we analyse the discussion across time, we observe a great variety of topics. Stream graph visualisation (Figure 9) provides a diachronic view of most frequent hashtags per month. The hashtags that are constantly present are very few: #Unibo for the university, #openconsiglio_bo related to the city council activity, #bambini (children) and #scuola (school), political parties and the prime minister #renzi and finally #viabilità (mobility). Then, contents change across time. In the first months, keywords related to social movements and strikes are more frequent: #omophobia in August and September 2013; #socialstrike, #sollevazione, #assedio and #occupymensa in October and November. Then, in 2014, discussion is principally political until the peak of May 2014 with the European elections.

[12] On these samples, geo-tagged tweets are practically inexistent.

Fig. 8: Word clouds of the terms included in tweets mentioning photos in tweets geo-tagged in Bologna between the 2nd and the 29th June 2014. General terms are excluded. Max 150 words shown.

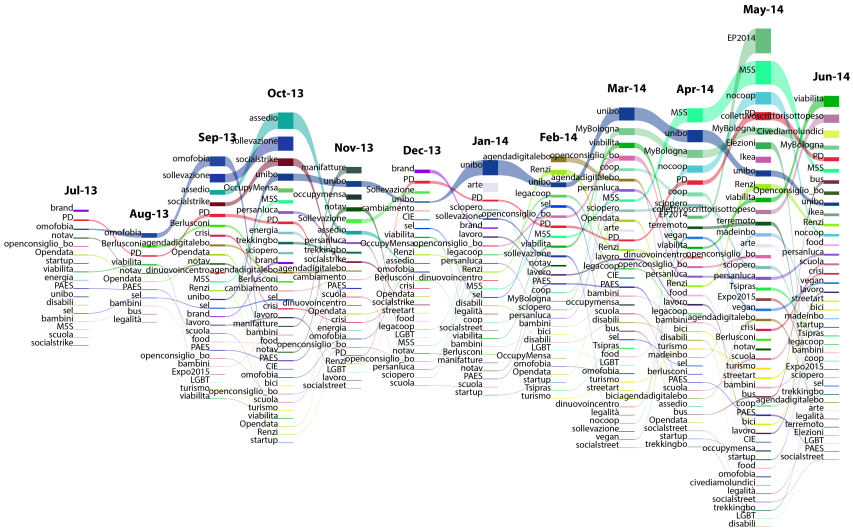


Fig. 9: Stream graph of top hashtags per month in tweets of Bologna's influencers (July 2013 – June 2014). Sport hashtags and "#Bologna" are excluded.

If we focus on aspects related to spatial analysis and urban planning, several insights can be useful to decision makers:

– Discussion about the city brand. In last months, Bologna council organised an important international competition for changing the brand of the city. Twitter has been a good terrain for the promotion of this project (in our corpus 140 tweets sent by 22 users mention "brand"). As an example, we use co-hashtag network analysis to have an overview of the discussion about brand. Through this graph (Figure 10), we can easily identify the principal themes related to the competition such as food, art, tourism and social.

Fig. 10: Co-hashtag analysis in tweets of Bologna's influencers mentioning "brand".

– Discussion about city politics. Several hashtags are related to city politics, notably #openconsiglio_bo and #iperbole.
– Discussion about energy. Energy is also a key topic especially through the hashtag #PAES (Patto dei Sindaci - Piani d'azione per l'energia sostenibile, Covenant of Mayors - Action plans for sustainable energy) that is present along all the period. Bioarchitecture and energy-sustainable housing are frequently discussed.
– Discussion about transports. It is interesting to know that main topics are bicycles (#bici, #ciclabile) and buses (#bus).

- Discussion about school. As said #scuola (school) and #bambini (children) are constantly present in the Twitter communication, yet these hashtags are mentioned in very varied contexts, from crimes to political declarations to the promotion of events.
- Discussion about enterprises of the metropolitan area. Several industries are mentioned, #coop and #coopservice, but also the boycott actions against Granarolo and the social demonstrations at Ikea.
- Discussion about the environment. Several hashtags are related to the environment, notably #bolognasostenibile, but also #ambiente (environment), #alberi (trees), #waterbologna. . .
- Discussion about the smart city. The impact of digital technologies on the urban life also appears: #BolognaSmartCity, #Smartcity, #agendadigitale, #bolognaagendadigitale.
- Discussion about San Luca, a church on the hills of Bologna. The crowdfunding campaign for the arcade of San Luca is diffused in several tweets such as "Cosa rappresenta il Portico di #SanLuca per #Bologna?Potete raccontarlo compilando un questionario online @Twiperbole `http://t.co/mZw6GUqK5z`" or "#SanLuca: parte la raccolta fondi per il portico `http://t.co/v7S4iY4cGC` 10 Ottobre il primo evento #Bologna #sostieniilportico @twiperbole". San Luca can be considered as one of the main symbols of the city.
- Discussion about change. Several hashtags related to change are used: #cambiamento, #cambiamotutto, #cambiareverso, #cambiaverso.
- Discussion about crisis. Hashtags like #crisi (crisis) and #austerity are often used. We have 362 tweets sent by 30 users that mention them. Through word frequency analysis of this corpus (Figure 11), we can identify the main content related to the crisis. The discussion generally involves harsh terms with words like "sollevazione", "assedio" and "ipocrisia" related to social demonstrations. The strong connection with Christmas is also interesting. Finally, we can also identify a positive attitude through some words: "ribaltiamolausterity", "anticrisi", "opportunita", "autoriduzione".

Fig. 11: Word clouds of tweets from Bologna influencers mentioning "crisi" and "austerity". The terms "crisi", "austerity" and "bologna" have been excluded from the representation.

Also European questions are touched upon in the discussions of influencers. Most of the mentions are related to the European elections (#EP2014, #Europee), yet there are also other mentions of Europe (#Europa) in other contexts such as a conference on the value of the arcade of Bologna in an European context and then some social movements such as #Civediamolundici. We can also note the strong attention for the visit of Alexis Tsipras. (RT @tsipras_eu: @LidiaSenra "we are going to build the Europe of workers bread and dignity" #AGEenEuropa #ELChangeEurope `http://t.co/NqZXx`...). What is interesting is that the most retweeted message of this corpus does not relate to European elections but to the possible effect of Renzi election on the European policy "RT @matteorenzi: Un risultato storico. Commosso e determinato adesso al lavoro per un'Italia che cambi l'Europa. Grazie #unoxuno. @pdnetwor...".

In the corpus, there are 1,919 tweets sent by 36 users that mention "euro*". We generate a co-hashtag network and a word cloud from this sample (Figure 12 and Figure 13). The first graph is very useful to identify different items of the discussion: the parties, links to the crisis and social demonstration, the Italian policy of Renzi (80euro) and also the presence of several tweets related to sport. Conversely, the word cloud helps us to underline the different weight of these items. After excluding contents related to sport, we see that political campaign "cominciatu" clearly emerges as the main content.

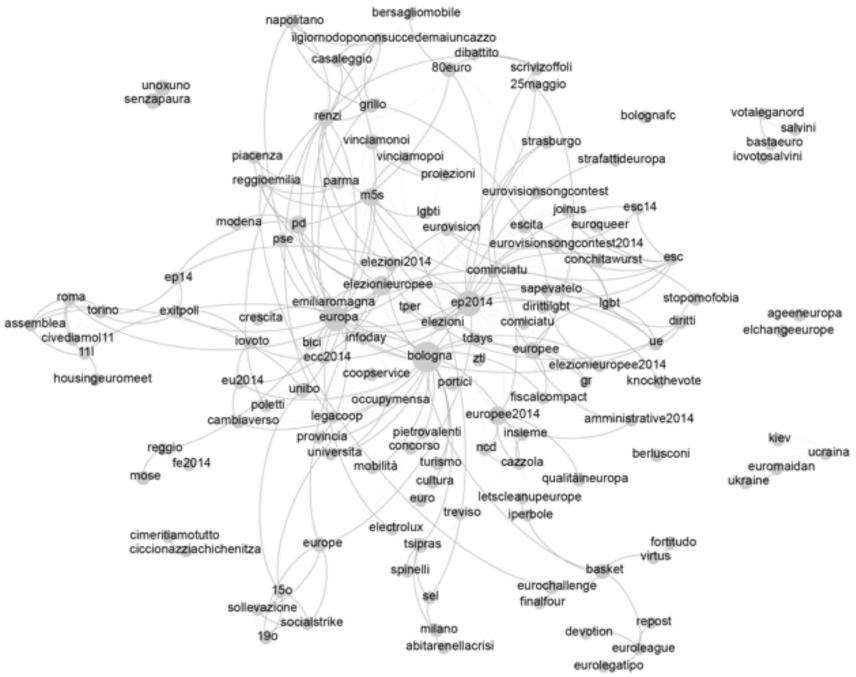


Fig. 12: Co-hashtag analysis in tweets of Bologna's influencers mentioning "euro" and "ep2014".

Given all that, Bologna's influencer's discussion is clearly centred on the life of the city. By building a good sample, decision makers can be updated about the key questions related

to their work. Through Twitter data, they can have a multiple viewpoint on their present and future actions and on the needs of citizens and they can follow them in real-time.

5 Conclusion

Through this case study, we add important insights about possible uses of Twitter data for public policy. As said, particular attention has been paid to methodological issues. We focus on the phases of the data collection and on the tagging process. As a final result, we could validate the interest of Twitter data for studying the city brand.

As regards to the selection of the data sample, first of all we can observe that methods based on the use of Web 2.0 data are grounded on a delicate balance between advantages and drawbacks of these data. On the one hand, these data promise to provide very new information about social life. On the other hand, the use of these data raises several methodological issues. In this project, we note two main issues: (a) the different uses of Twitter, notably its ephemeral and expressive use (Bruns & Moe, 2014). Yet, if we consider tweets mentioning a city, the noise related to fun and emotions does not prevent to identify key topics related to the city. At the local level, the problem of the different uses of Twitter can be effectively avoided by analysing a selection of tweets sent by a restricted number of users identified as city's influencers. This project clearly demonstrates that a pertinent choice of Twitter accounts will reduce the noise found in the first two samples. (b) The small amount of geo-tagged tweets. Actually, geo-tagged tweets constitute just a very small percentage of global Twitter activity (Gerlitz & Rieder, 2013). Even if this is surely a very big issue for the selection of data, two observations may be made on the basis of the experience of this project. First, if geo-tagged tweets are not so useful as absolute values they can be used in a comparative way. In the first method, we see that the analysis of the over-quotation distribution of city's mentions can be considered a good indicator of the international recognition of a city. Moreover, in the second method, we see that the content of geo-tagged tweets is not so related to space, so geo-tagged tweets are not necessarily the good entry for studying spatial questions. We agree with Crampton et al (2013) on the importance of going beyond geotagging.

As regards to how a big data sample can be tagged, we can note that tagging data was necessary in order to carry out content analysis of tweets and it is not evident to choose between quantitative and qualitative methods. Considering the size of Twitter datasets, automatic tagging procedures are surely more desirable because less time-consuming. Yet, the use of Twitter data raises the opposition between the necessity of automatic procedures of tagging and the high possibility of tagging errors. Besides the initial fears, results of automatic techniques for analysing tweets mentioning a city are encouraging. The analysis of most frequent hashtags can be a very efficient technique for identifying public events happening in the city (by week). Moreover, word clouds of geo-tagged tweets can have two main objectives: (1) finding errors due to homonymy; (2) identifying the main elements of the image of a city in different countries. So, we can say that automatic methods can be a profitable strategy, even if a preliminary effort of disambiguation is necessary in order to anticipate main possible errors (other meanings of the name of the city, other places with the same name).

Then, more qualitative methods can provide a more complete overview of the themes discussed about the city. We propose to manually categorise tweets and to represent their distribution in categories through a treemap. This visualisation tool can be very useful for decision makers that can have a general point of view on topics related to the city and that can quickly identify hot topics that affect the image of the city at national and international level. This technique can be partially automatized by building a database of correspondences between hashtags and categories. Yet, it is important to underline that in content analysis of tweets, even when we are using an automatic technique, qualitative interpretative work is equally necessary. The style usually employed for writing tweets can be not so immediate for the researcher or the decision maker.

As a conclusion, we can note that, before the emergence of the Web 2.0, decision-makers could obtain insights about public opinion principally by media or by more expensive techniques such as interviews, focus groups and public surveys. Obviously, Twitter users cannot represent statistically all population of a city. Yet, tweets are expected to give innovative insights on social life useful for public policy. This project shows that, through Twitter data, it is possible to identify new and more varied sources of information about the brand of the city such as the interests of people living in the city or speaking about the city, the attractiveness of some special events happening in the city, or the connections of the city at the international level. These sources can be varied and uncontrolled as in the first method and they can provide quantitative overall measures of the international recognition of the city. Or, by selecting a specific sample of Twitter accounts (called influencers), decision-markers can have a real-time and on-going feedback about the key questions related to the city. Decision-makers can distinguish permanent topics and emerging topics by following the general use of hashtags and they can focus on specific hashtags related to the political agenda. Yet, based on these first positive insights, future research has to been carried out in order to validate these results on a longer period.

Fig. 13: Word cloud of tweets of Bologna's influencers mentioning "'euro" and "ep2014" excluding tweets containing "euroleague" et "eurochallenge". In the reprentation the following words have been excluded: Bologna, Emilia, Euro*, EP.

Bibliography

[1] Borra, E. & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66, 3, 262-278.

[2] Bruns, A. & Moe, H. (2014). Structural Layers of Communication on Twitter. In Weller, K. et al. (Eds) *Twitter and Society* (pp. 15-28). New York: Peter Lang.

[3] Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data'and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40, 2, 130-139.

[4] Davis Jr., C. A., Pappa, G. L., de Oliveira, D. R. R. & de L. Arcanjo, F. (2011), Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*, 15: 735–751. DOI: 10.1111/j.1467-9671.2011.01297.x

[5] Gerlitz C. & Rieder B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal*, 16, 2.

[6] Sakaki T. , Okazaki M. & Matsuo Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web*, April 26-30, Raleigh, North Carolina, USA.

Building a recommender system using multilingual multiscript tweets

Ritesh Shah[1,2], Christian Boitet[1], and Pushpak Bhattacharyya[2]

[1] GETALP-LIG, Université Grenoble-Alpes,
Grenoble, France
ritesh.shah,christian.boitet@imag.fr
[2] CFILT, Indian Institute of Technology,
Mumbai, India
pb@cse.iitb.ac.in

Abstract. We describe an ongoing work aimed at building a Recommender System by applying some innovative Natural Language Processing (NLP) techniques on multilingual, multiscript Twitter data (tweets). Large quantities of tweets in some Indian languages, French and English will be used for processing. Multilingual tweets often exhibit some degree of code-mixing, style variations and disfluencies. Therefore, we rely on an interlingual content extraction approach and apply pertinent NLP techniques on tweets in a modular way. The modules of the multilingual Recommender System are preprocessing, multilingual morphological and presyntactic analysis, named entities identification, interlingual lexical disambiguation (using UWs[3], UNL[4] interlingual lexemes), content extraction (with reference to a very simple ontology or knowledge base) and polarity determination.

Keywords: SRecom, recommender system, twitter, tweet, multilingual, multiscript, code-mixing, UNL, UW, NLP, French, Indian languages.

1 Introduction

Recommender systems (SRecom(s)) are useful when built upon services like Twitter where social information exchanges are characterised by both speed and volume. Recommendations based on such large amounts of information reportedly enable users to make quick and better decisions. In mining this information, nevertheless the magnitude and the inherent diverse characteristics of tweets may pose unique problems. Also, the design of SRecom(s) needs to be driven by specific tasks, information needs, and item domains [6]. It is then essential to identify, accommodate and address these concerns while building a SRecom based on tweets. Presently, the approaches mostly used in building SRecom(s) are collaborative filtering [20] and content-based approaches [11]. Furthermore, the strengths and limitations of each of these approaches have also led to hybrid methodologies, such as those of [2, 3, 13].

[3] Universal Words
[4] Universal Networking Language (http://www.undl.org)

In this paper, we propose an implementation of a *multilingual Recommender System* (*m_SRecom*) based on tweets while keeping in mind the relevant concerns. Our motivations in building a m_SRecom are as follows.

1. A significant potential demand for tourist-oriented recommendations as exhibited by numerous tweets.
2. Availability of NLP resources and competences for several languages in our lab including some Indian languages, French and English.
3. Availability of large Twitter datasets (through the Twitter API) concerning the same sub-domain of the tourism domain in several languages.

In our work, we plan to apply NLP techniques on multilingual tweets and to make use of an interlingual content extraction approach. We select some Indian languages, French and English as the linguistic base and confine our m_SRecom to provide tourism related recommendations. Our approach demands careful investigation of the underlying multilingual issues and, more importantly allows us to assess the applicability of NLP in implementing an effective m_SRecom. Altogether, we aim at handling several scientific and technical challenges set forth by the micro-blog characteristics, the specific recommendation needs and the proposed multilinguality.

In the subsequent sections, we begin with a short background on Twitter and tweets. This is followed by a description of the tweet collection methodology and a preliminary study related to code-mixing and geo-information in tweets. The concluding section contains a description of each of the NLP modules proposed for the m_SRecom.

2 Background

2.1 Twitter

Twitter is a publicly available rich source of social data in the form of short status updates called *tweets*. Twitter produces 45 Giga[5](G) tweets per quarter as of early 2015. This roughly translates to 500 Mega[6](M) tweets per day. Since its public launch in 2006, the increase in tweets produced worldwide has been phenomenal as shown in Table 1. It is notable that regional distribution of tweets is difficult to estimate because the tweets are usually devoid of reliable geotagged information.

Twitter provides access to tweet data primarily through the REST[7] APIs and the Streaming APIs. The REST APIs are well suited for running searches against the index of recent tweets, whereas a Streaming API allows near real-time access to a subset of tweets by random sampling, for instance. Access requests from the developers are authorised through an OAuth[8] service. The API calls are limited per API method and also depend on the authentication mechanism explained in the Twitter documentation.

[5] 1 Giga = 10^9
[6] 1 Mega = 10^6
[7] Representational State Transfer
[8] Open standard for Authorization

Table 1: Growth rate: Number of tweets generated worldwide.

Year	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
# tweets per quarter	60K	400K	100M	225M	5.85G	18G	30.6G	36G	45G
# tweets per month	20K	133.3K	33.3M	75M	1.95G	6G	10.2G	12G	15G
# tweets per day	667	4.44K	1.11M	2.5M	65M	200M	340M	400M	500M

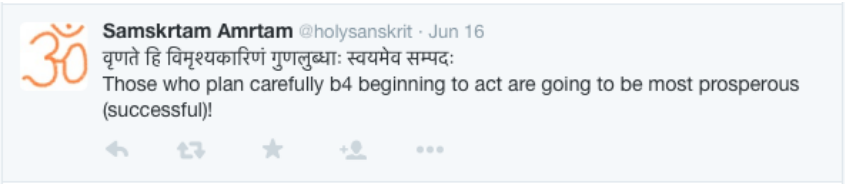


Fig. 1: An example of a tweet. Source: https://twitter.com

2.2 Tweets

A 'Tweet' is any message posted to Twitter which may contain images, videos, links and up to 140 characters of text.

Underlying a tweet is a rich set of metadata (typically 5 KB in size) in the form of embedded attribute-value pairs. It is essential to understand these metadata to effectively use the Twitter APIs. Some fields are briefly described as follows.

id:	integer representation of unique identifier for this tweet
text:	contains the actual UTF-8 text of the status update
retweet_count:	indicates the number of times this tweet has been retweeted
retweeted_status:	contains a representation of the original tweet
place:	indicates that the tweet is associated with a place
geo_enabled:	indicates whether geo-tagging is enabled by the user
lang:	indicates user's self-declared user interface language

3 Preliminary experiments

3.1 Data collection

We use an interactive Python environment for programming and development, and the Twitter search API to extract tweets. The search API (non-Streaming API) allows the developer to obtain a maximum of 1.73M tweets/day through the *Application-user authentication* (AuA) and a maximum of 4.32M tweets/day through the *Application-only authentication* (AoA).

These API constraints necessitate deployment of a large number of computer resources if the goal is to collect all 500M tweets/day produced worldwide as in Table 1. More than 115 distinctly authenticated processes in parallel would be needed to run than is theoretically required.

But, if the goal is only to collect tweet samples based on search queries with filters (for instance, language or location related filters) then the task seems to be manageable. For instance, [10] reports that during the electoral polls of 2014 in India, *election* related tweets were produced in the range of 0.54M/day and 0.82M/day from April 1 to May 12.

Even if we assume that the production of *tourism* related tweets per day in India is ten times more, they can be collected with less than 5 distinctly authenticated processes using suitable APIs. The search API returns a collection of tweets corresponding to the requested query and the specified query filters.

To begin with, we looked at some multilingual tweets obtained by submitting different queries through the Twitter web interface. We then selected the queries that yielded the most suitable results and used them in programming scripts to access the search API. As a preliminary study, in the following sub-sections we describe two experimentations on the tweet results obtained by the search query Q1.

Q1: 'lang:hi OR ("indian music" OR "indian festival" OR "indian restaurant") '

We use the search API under the AuA mode to acquire a sample of 100K tweets using Q1. The result is a set of all tweets filtered on the field 'lang:hi' (with Hindi as the user-indicated language) or tweets matching either of the 3 consecutive terms.

3.2 Code-mixing: a preliminary study

In the context of tweets, the text is often full of spam, incorrectly spelt words, disfluencies, slang words and other anomalies, as shown by [14]. Moreover, even though these texts are quite short, they exhibit some degree of code-mixing. So, we tried to estimate the proportion of code-mixed text in the 100K tweet sample.

The step-wise formulation of the experiment is as follows.

1. For a query Q, we obtain a set of tweet responses denoted by $T(Q) = \{R_1, R_2, R_3, \ldots\}$. Each R_i is a response in JSON[9] format representing metadata of each tweet.
2. For each R_i, we retrieve the values from the 'text' field containing the user's tweet message, s_i. Let $S(Q) = \{s_1, s_2, s_3, \ldots\}$ be a set of all tweet message(s) obtained from query Q.
3. For each s_i, we collect whitespace-separated tokens (words) in a set denoted by $W(Q) = \{w_1, w_2, w_3, \ldots\}$.
4. We now use regular expressions (regex) to group all the words in $W(Q)$ into the following collections.

[9] JavaScript Object Notation

ALL:	The set W(Q) containing all the words.
FILTERED:	Set of words excluding hashtags, screennames, URLs and 'RT' (retweet) words. Words that do not begin with '@' or '#' or 'http'. The word 'RT' is also excluded.
ALPHANUM:	Set of words that only match the alphanumeric characters, regex:'[A-Za-z0-9]+'.
NON ALPHANUM:	Set of words that only match the non-alphanumeric characters.
HASH TAGS:	Set of hashtags. Words that begin with '#'
SCREEN NAMES:	Set of screennames. Words that begin with '@'
MIXED:	Contains words having atleast one alphanumeric character and one non-alphanumeric character.

A large fraction of tweets in the 100K sample was in Devanagari script because of the 'lang:hi' attribute in the query. So, in Table 2 we calculate code-mixing as a ratio of *tokens in ALPHANUM* to the *tokens in FILTERED*. In these data, the average code-mixing percentage at the word-level over 100K tweets is 3.28%. That is certainly a very small proportion, but it nevertheless corresponds to a very large number of tweets everyday (approx. 15K tweets/day assuming a 500K tweets/day originating from India, [10]) and they may presumably contain very interesting information (produced by visitors, for example).

3.3 Geo-information: a preliminary study

Third-party information services built on social media platforms often provide statistics that help to perspectivise and research social media behaviour. A similar study on worldwide distribution of Twitter users by Statista[10] shows an increase of Twitter users from 29% in 2012 to 35% in 2015, especially in the Asia-Pacific region. Geo-information contained within tweets is one way of determining the origins of tweets and Twitter users.

However, this information is unreliable and needs investigation as undertaken in a recent work by [12]. We are interested in acquiring multilingual tweets based on related scripts or based on originating locations. So, we do a basic check of the extent of geo-information present in the tweets using the 100K tweet sample. The location information, other than for geo-enabled tweets, are updated by users and could be unreliable. We parse the tweet JSON structure to collect non-empty values from the located related attributes in the tweet.

The numbers in Table 3 indicate the presence of location information in the tweets. From the 100K sample, we observe that a total of 48458 (48.5%) tweets are geo-enabled.

4 Modules and methodology

There are 6 main modules in our proposed pipeline.

1. Twitter data: Preprocessing, Extraction, Analysis and Storage (PEAS)

[10] http://www.statista.com/statistics/303684/regional-twitter-user-distribution/

Table 2: (a) Proportion of word-level code-mixing over 100K tweets. (b) Rate limited API allows a maximum of 18000 tweets per request

	Collection Types						
#tweets	**ALL**	**FILTERED**	**ALPHA NUM**	**HASH TAGS**	**SCREEN NAMES**	**MIXED**	**Code mixing %**
16000	309384	262594	8471	7376	19411	6775	**3.23**
16000	312700	267851	6888	5688	19646	5790	**2.57**
16000	317013	272523	7024	4967	19355	5908	**2.58**
16000	310550	267945	10761	4754	18910	6643	**4.02**
16000	304331	259485	8264	4862	19867	6141	**3.18**
16000	293409	250810	8270	4735	19274	5735	**3.3**
4000	72974	62280	2528	1457	4313	1573	**4.06**

Table 3: Distribution of non-empty geo-related fields.

#tweets	place	coordinates	location	time_zone	geo_enabled
16000	172	19	10175	4491	7444
16000	244	64	10845	5546	8028
16000	204	74	10795	5957	8234
16000	223	66	10564	6359	7900
16000	311	51	10636	6637	7898
16000	256	54	10664	7013	7481
4000	42	7	2212	1549	1473

2. Morphology-based Processes and Presyntactic Analysis (MPPA)
3. Named Entities identification (NEi)
4. Interlingual Lexical Disambiguation (WSD)
5. Content Extraction (CE)
6. Polarity Determination (PD)

4.1 Twitter data: Preprocessing, Extraction, Analysis and Storage (PEAS)

Our preliminary observations on code-mixing and plans in doing a detailed study of the same are supported by the following published work. [5] discuss code-mixing rules and constraints in English-Hindi mixed language data, while contrasting code-mixing in tweets as opposed to free-form text. [17] shows increasing trends of code-mixing in the past decades.

As seen in the preliminary observations, an average of more than 3% of the Hindi tweets are code-mixed that can possibly be significant for further analysis. An example of code-mixed tweet in Marathi (another Indian language) sharing the same script as Hindi can be seen in Fig. 2.

As further necessary steps in processing code-mixed data, we plan to:

1. build or assemble tools for collecting the largest possible set of tweets in the concerned languages,
2. perform multiple and elaborate analysis of their languages and scripts,
3. normalize them to a standard encoding (UTF-8) and standard scripts.

The preprocessed data will then be stored in a multilingual corpus management system such as SECTra[11].

4.2 Morphology-based Processes and Presyntactic Analysis (MPPA)

The goal is to build compatible morphological analysers (producing the same kind of output) for all languages, and to use the results to identify named entities (names of places, events,

[11] See [9], [1], [15].

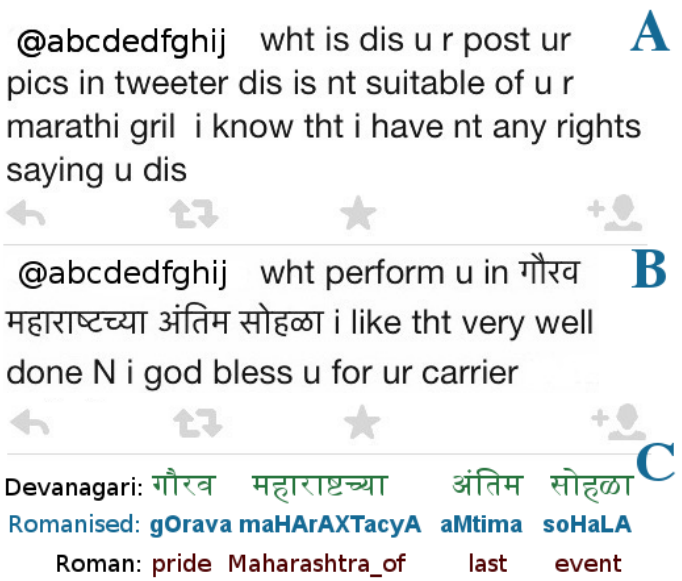


Fig. 2: Example of a code-mixed tweet
(A) tweet 1: with typos and contracted terms in English (Roman script)
(B) tweet 2: in English (Roman script) and Marathi (Devanagari script)
(C) Marathi transliteration: « gOrava maHArAXTacyA aMtima soHaLA »
« pride of_Maharashtra last event »

people, etc.) and finally to annotate the simple or compound lexemes found by their possible word senses, expressed as UNL interlingual lexemes (UWs). An interesting and quite new perspective is the possibility to build a unique morphological analyzer for the set of languages (rather, sublanguages) and for the various writing conventions at hand. For this, we plan to use a new version of ATEF, a tool built on an extended NDFST model by [4][12].

4.3 Named Entities identification (NEi)

Identification of named entities shall follow a simple syntactic analysis based on "local grammars" or "local automata", where the goal is only to find small linguistic fragments, or "chunks", such as elementary noun phrases (that are often named entities or terminological units), and then structure them in small multilevel structures or small UNL graphs [19].

A favorable circumstance in the chosen domain (tourism) is that quite a large number of named entities are available and updated in external free resources (persons, places and above all, various types of events).

4.4 Interlingual Lexical Disambiguation (WSD)

These processes usually involve numerical or statistical computation, sometimes even a degree of Machine Learning. They concern lexical disambiguation (WSD, leading to score assignment

[12] See also [8].

to possible UWs), semantic disambiguation (at the level of small groups/phrases), and typically use conceptual vectors, ant algorithms, and similarity or semi-distance computations [18].

4.5 Content Extraction (CE) and Polarity Determination (PD)

[16] conclude that building SRecom(s) using ontologies and knowledge bases (KBs) reduce several specific problems related to:

1. inter-operability of resources and homogeneity of representation,
2. dynamic contexualization of user preferences,
3. semantic extension of descriptions using other contextual factors.

Following them, and building on previous work in the OMNIA project [7], we will use Semantic Web technology for defining the "knowledge", that is, we will use one or more small KBs or small domain specific ontologies. For the proposed m_SRecom, the content extraction and polarity determination operation has to be done relative to each knowledge base (e.g., a relational database, or an ontology) in which one wants to index each text (tweet in this case). It will work in a language-independent way, using the disambiguated word senses (UWs) and their scores, and relating to each KB by using a UW-KB alignment.

Acknowledgments

We would like to thank UJF (Université Joseph Fourier), IRD (Institut de Recherche pour le Développement) and IITB, (Indian Institute of Technology, Mumbai) for supporting this research.

Bibliography

[1] Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. The PAPILLON project: Cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In *"Proceedings of the 2nd Workshop on NLP and XML at COLING'02*, pages 93–96, Taipei, 2002.

[2] Svetlin Bostandjiev, John O Donovan, and Tobias Höllerer. TasteWeights : A Visual Interactive Hybrid Recommender System. In *"Proceedings of the 6th ACM conference on Recommender systems - RecSys '12"*, pages 35–42, 2012.

[3] Robin Burke. Hybrid web recommender systems. *The Adaptive Web*, pages 377–408, 2007.

[4] Jacques Chauché. The ATEF and CETA systems. In David Hays, editor, *American Journal of Computational Linguistics (2, microfiche 17)*, chapter 1, pages 21–40. Association for Computational Linguistics, 1975.

[5] Anik Dey and Pascale Fung. A Hindi-English Code-Switching Corpus Code-Switching in Indian Culture. In *"Proceedings of the 9th Language Resources and Evaluation Conference"*, pages 2410–2413, 2014.

[6] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative Filtering Recommender Systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2010.

[7] Achille Falaise, David Rouquet, Didier Schwab, Hervé Blanchon, and Christian Boitet. Ontology driven content extraction using interlingual annotation of texts in the OMNIA project. In *"Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING'10"*, pages 52–60, Beijing, August 2010.

[8] Jean-Philippe Guilbaud, Christian Boitet, and Vincent Berment. Un analyseur morphologique étendu de l ' allemand traitant les formes verbales à particule séparée. In *"Proceedings of TALN-RÉCITAL"*, volume 9, pages 755–763, Les Sables d'Olonne, June 2013.

[9] Cong-Phap Huynh, Christian Boitet, and Hervé Blanchon. SECTra_w.1: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. In *"Proceedings of the Language Resources and Evaluation Conference"*, volume 8, pages 27–31, 2008.

[10] Raheel Khursheed. India's 2014 #TwitterElection. `http://blog.twitter.com/2014/indias-2014-twitterelection`, November 2008.

[11] John Bruntse Larsen. *Content-based Recommender Systems.* PhD thesis, Technical University of Denmark, 2013.

[12] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *"ACM Transactions on Intelligent Systems and Technology"*, 5(3):47–69, 2014.

[13] Prem Melville, Raymond J Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *"Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)"*, pages 187–192, Edmonton, July 2002.

[14] Subhabrata Mukherjee, Akshat Malu, A R Balamurali, and Pushpak Bhattacharyya. TwiSent: A Multistage System for Analyzing Sentiment. In *"Proceedings of the Conference on Information and Knowledge Management (CIKM)"*, pages 2531–2534, 2012.

[15] Hong-Thai Nguyen, Christian Boitet, and Gilles Sérasset. PIVAX, an online contributive lexical database for heterogeneous MT systems using a lexical pivot. In *"Proceedings of the International Symposium on Natural Language Processing"*, Bangkok, 2007.

[16] Eduardo Peis, JM Morales del Castillo, and JA Delgado-López. Semantic recommender systems. An analysis of the state of affairs. *Hipertext.net, Spain*, 6:1–5, 2008.

[17] A. Si. A diachronic investigation of Hindi-English code-switching, using Bollywood film scripts. *"International Journal of Bilingualism"*, 15(4):388–407, 2011.

[18] Andon Tchechmedjiev, Jérôme Goulian, Didier Schwab, and Gilles Sérasset. Parameter estimation under uncertainty with Simulated Annealing applied to an ant colony based probabilistic WSD algorithm. In *"Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology"*, pages 109–124, 2012.

[19] Bernard Vauquois and Christian Boitet. Automated translation at Grenoble University. *"Computational Linguistics"*, 11(1):28–36, 1985.

[20] Xiwang Yang, Yang Guo, Yong Liu, and Harald Steck. A survey of collaborative filtering based social recommender systems. *Computer Communications*, 41:1–10, 2014.

Colors of the World Cup: visualizations of images shared on Twitter during the 2014 World Cup

Fábio Goveia, Lia Carreira, Lucas Cypriano, Tasso Gasparini, Johanna Honorato, Veronica Haacke, and Willian Lopes

Department of Communications
University of Espírito Santo
Espírito Santo, Brazil
fabiogv@gmail.com,liacarreira@gmail.com

Abstract. The 2014 World Cup, held in Brazil from June 12 to July 13, had great impact on social networking sites. Seeking to better understand content sharing flow and to explore methodological possibilities for studies with large volumes of images data, the Image and Cyberculture Studies Lab (Labic), of the Federal University of Espírito Santo, created the project Colors of the World Cup - a project that presents interactive image visualizations of contents shared on Twitter during the football tournament. Throughout this paper, we will present the project's methodological process and propose a brief analysis of these images as a way of documenting the social and cultural movements and events that appeared on Twitter during the World Cup.

Keywords: Image, Visualization, Twitter, Color

1 Introduction

The large amount of information produced by each individual through mobile devices can be considered one of the most significant marks of contemporary society. On the one hand we are producing these data, on the other hand we are also sharing, circulating and consuming constantly these information. Digital social networks are among the most commonly used ways to consume and circulate personal and public news. These platforms have become privileged interaction zones, with more and more exchanges of information and experiences.

This new universe opens up numerous opportunities for knowledge and relationships - that once seemed nearly impossible to happen – and also makes it possible to develop research analyzing these interactions and weaving more precise understandings of the social behavior of individuals.

In this scenario new experiences among users emerge and important tools of activism and change get created. In addition to the significant role of social networks in the political field, it should be also be highlighted the significant amount of information produced in relation to other large social events (sports, cultural, artistic). Data from these social networks related to these events can contribute significantly to different studies, such as consumer behavior.

From this overview, the analysis of the ways in which such content spread through the networks is necessary for a better understanding of how these social movements develop and behave, in and out of these digital environments. Furthermore this analysis can help to identify potential patterns or future occurrences. Thus the big data studies appear as a field of study that has been highlighted in recent years, especially due to the expansion of this large amount of information circulating in the environment of digital social networking, and that has been used for several studies, as the one presented here.

Seeking investigate new methodological frameworks, the Laboratory of Research on Image and Cyberculture (Labic), started the World Cup Colors project, which sought to analyze the production of images that occur in social digital networks during the 2014 World Cup held in Brazil. The event took place between June 12 and July 13, 2014, with a great impact on social networks. On this occasion, we have seen emerging a large number of images that were critical or made fun of the event, giving way later to a more favorable outlook. This change was attributed to the network called "Boleira" that is, the network of those fan profiles that accompany the real-time games, the backstage and the memes about the World Cup. They significantly contributed to the expansion of the online debate to a more positive and less critical bias.

Before the Cup a strong tension was expressed by various popular protests in several Brazilian cities. The first analysis of the images on social networks revealed that this tension ultimately was displaced by the accounts of the football fans which dominated the sharing of images on Twitter. No other perspective occupied the media spaces network in the same way as the football fans profiles. At one point, even some activists that were complaining about the FIFA impositions to the country, and corruptions with the event, started communicating about the games and became part of this football narrative.

In the following, we present the methods that guided the development of this research, as well as web application development, launched on October 6, 2014 in order to provide viewing modes for large sets of images collected during the World Cup.

2 Methodology

Manovich and Hochman (2013); Manovich et. al. (2011); Crandall et. al. (2009 e 2015) are some of the authors that work in this area, from social networks data mining to development of big data image visualizations. The challenges proposed by them were also part of this research. In this research were collected images shared on Twitter. In this case, data was collected through World Cup[1] related terms such as specific keywords, hashflags[2] of each country, terms of each match (for example, #BRAxGER used for the game between Brazil and Germany), and the names of the major soccer players, such as "Neymar," "Messi" and

[1] Terms were used that made reference to the main event, as "world cup", "copa2014", "brasil2014", "worldcup"; the side events directly related to WordCup as "Funfest", "wordcup opening"; and the organizations responsible for the event, such as "fifa" and "cbf".

[2] Hashflags were a special kind of hashtag used on Twitter during the World Cup. Were hashtags related to the countries participating in the World Cup.

"Podolski". Here, the aim was to collect, from June 12 to July 13 of 2014, the largest number of tweets related to the event, using precise terms and related keywords.

The tweets published during this period were collected from a script called Marcus, developed in Python by Labic in partnership with the computer scientist André Panisson. This tool searches the tweets containing text terms previously selected and stored in a remote server database (MongoDB[3]). After this first mining, every 15 minutes of collect by Marcus script another script called Crawler[4], also developed by Labia, accessed the database, captured the tweets of that period containing links (possible images), eliminates those images links resulting from external sites, considering in this way just the images that was posted directly in the tweet. After this the Marcus script selected the 100[5] most shared images during the selected period - this mean, the images links most repeated during the selected period, in this case each 15 minutes. Thus the script accessed each link collected, saved the image attached and generated a .csv table listing the link of the image and the saved file for further analysis.

With the image dataset ready in hand, it was then possible to proceed with the data processing and image analysis, done in two simple steps: first the extraction of chromatic data and, second, image comparison in order to eliminate repeated ones. Therefore, at the end of each day's data mining process, a third script also developed by the lab called AISI (Automatic Identifier of Similar Images) swept the dataset, extracted and compared the image's colors metadata (such as hue, brightness and saturation, as well as each image's histogram), and identified those that were similar to each other, avoiding repetition. When similarities were identified, these similar images are considered as single ones, adding their frequencies (i.e., the amount of time they were posted) as to give them their correct weight in the dataset. Without this procedure, each one of these similar images, published in different tweets, wouldn't have the impact they actually have on these networks through online sharing - a very important characteristic of social media content.

With this extraction method, we obtained a total of 30 million tweets, of which 2 million contained links with images, and a total of 42,000 images posted directly on Twitter. Among these, there were 17,000 unique images that were then used to create the visualizations. With these data, it was possible to launch a web application called Colors Of the World Cup that facilitates reading this diversity of content by organizing it according to previously collected data, such as their chromatic characteristics and frequency of sharing.

The online interface of the application was developed using the JavaScript programming language, and the D3[6] library. From the .csv files generated by the Crawler and AISI, the

[3] Open source database, developed in C ++ programming language. The system generates an information store in JSON documents, a data file type that works independent of any specific programming languages. More information available at: `http://www.mongodb.org/`

[4] Script developed in Java, that accesses the selected links and downloads of these the images found. Available code: `https://github.com/ufeslabic/crawler`.

[5] For to do the extraction of all images published would require a long time, it would preclude the 15-minute time window.

[6] D3 (Data Driven Documents) is a JavaScript library designed to handle large data files for online visualizations. More information available at: `http://d3js.org/`.

engine retrieves information to generate the views and display them in the web application. With this data four different visualizations are generated: chromatic calendar[7], chromatic timeline[8], chromatic mosaic[9] and images mosaic[10].

The chromatic calendar (Figure 1) brings daily charts, containing circles as representations of each image from its predominant color. In this view they are counted only the shares obtained by the image in each day. The circle (that represent a unique image) position on the X axis is defined by the predominant hue of the image that its represents, while the position in the Y axis is defined according to the number of shares that the image obtained during that period; the radius of each circle brings the total number of shares that the image accumulated over time; and the color of the circle is defined by the predominant hue of the image. Hovering the mouse over each circle shows the image, and click in the view starts an animation that shows the behavior (sharing number) of images throughout the day in 15-minute intervals.

The chromatic timeline (Figure 2) is similar to chromatic calendar, but now shows on the same graph the total volume of the images during all days of the extraction period, and with this allows to be observed the behavior of each image along all the days. Same as the chromatic calendar the X axis position and the circle color is defined by the predominant hue of the image that its represents. But now the Y axis position is defined according to the number of shares that the image obtained during all one day, and the radius of each circle brings the total number of shares that the image accumulated over the days. Clicking the graph shows an animation with the behavior of images in each day. You can also pause or fast forward the animation to view the images more closely every day, and skip to specific days.

The chromatic mosaic (Figure 3) is a more straightforward view that brings the images represented by smaller circles with their respective predominant color and organized by its color hue. And hovering the mouse over each circle shows the original image.

Finally, the image mosaic (Figure 4) is the last visualization, in this is displayed the own images organized by their predominant hue color value, with a option of zoom in and out to have a more overview of all images or a more close visualization for a specific image. It is important to note that this is the only visualization of the application that uses the stored images[11] from the dataset, while the others visualizations access the original link of the image to show those in real time. This means that images in these others views will gradually fail to appear, since their source links may in time cease to exist. Thus, these views reflect the vital movements in networks.

Likewise, downloading and storing the images is also a way to keep these vivid images as memories of a movement or major event. Working with these large datasets is to work with

[7] Available at: `http://labic.net/coresdacopa/calendariocromatico/`

[8] Available at: `http://labic.net/coresdacopa/timelinecromatica/`

[9] Available at: `http://labic.net/coresdacopa/mosaicocromatico/`

[10] Available at: `http://labic.net/coresdacopa/mosaicoimagens/`

[11] Because of the large size of the image and to possibility zoom the view, the ZoomIt tool, developed by Microsoft for viewing very large size images on the Internet has been used. More information available at: `http://zoom.it/`.

their transitions and movements over a period of time, and create views is one way to realize these processes and make them visible.

3 Analysis

The Colors of the World Cup visualizations aims to highlight certain features of the shared content on the network during the 2014 World Cup. Through them, it is possible to highlight, from a wide range of content, key aspects about the way of images sharing in social network during the collection period. By analyzing these views, an intense chromatic variation of the images can be observed. A large volume of green tone images is noticeable, partly because the green color of the grass (since most of the photos with this predominant color tone are scenes that happened on the field) and partly because the presence of some Brazilians flags. There is also a considerable number of images in blue shades composed of uniform images such as the Argentina selection uniform with predominant blue colors, and a large number flags and advertising materials. In contrast, the more yellowish region, there is a large volume of images of materials related to the Brazilian national team, notably the yellow shirt traditionally used in games.

In the area with more red tones, we see mostly pictures of people, due to skin tones approaching beige and red. There is still a small part with pink predominant tone images, composed largely of images published with filters (like the ones used in the Instagram application), which pull the image tone to pink.

Beyond this evident chromatic variation, shared oscillations by day can be observed during the World Cup. This shows that the impact of the event on the networks did not occur in a homogeneous way. On the contrary, in a single day there was a very large amount of images with radically different levels in popularity on the network. The high frequency of an image in a day did not imply a similar level of popularity the next day, as happened with the eagle image (Figure 5) with the US flag on 22 June12, being the third most shared image of the entire Cup, obtaining total 22,098 posts just on the first day of sharing. This eagle image represented the discussions taking place on Twitter during the US game with Portugal, highlighting the broad participation of the North American and other supporters throughout this day. This demonstrates, therefore, that the image sharing modes during the World Cup were related to the specificity of the events of each day.

The most shared image throughout the period appeared on the network a few days before the end of the World Cup, on 5 July. It is an image of Neymar (famous soccer player) (Figure 6) kissing the ball, with a yellow mount in the background. The image was shared too by player David Luiz after Neymar suffered a spinal injury and cumulated a total number of 38,490 shares. This represents the support of fans and players in relation to their concerns for Neymar and was a very present image from the days of the event, showing the sentimental side of the players. Similar images in which players appear crying or celebrating a victory or defeat reinforce the emotional tone that the shared images had during the World Cup.

Thus, these images underline the existence of anemotional outlook: images shared by the network reflect the feelings of social networking users in relation to the championship and the participating teams. Images with scenes of the competition, as photo of players

being emotional in the field, were used to express the feelings of users. This is most clear in the images most shared between the 8th and 9th of July, which brought scenes of sadness resulting from the defeat of Brazil to Germany by 7-1, and in the second shared image (Figure 7) between the 12th and 13th June.

The wordcup shared images also have strong memetic content, ie, were created to become memes (images with the potential to multiply and propagate quickly, and can be given new meanings in the sharing process). These memes dominated the images on the networks that have formed around topics related to the World Cup, especially during the event first week, and remained largely shared together with pictures of other events that drew public attention (such as the elimination of teams considered "strong" or the Brazilian defeat to Germany, for example).

Thus, these visualizations made possible a rich analysis with the identification among other characteristics, of the existence of a chromatic rhythm. The images pulses and dances, even if at times, out of step and rarely able to maintain a constant frequency. This rhythm is symptomatic of the incessant and extremely fast content sharing process, characteristic of our time, that makes images suddenly appear, rise and quickly fall, preventing them to remain on the spotlight for too long. Thus, we can say that these social network images compete with each other for a spot to be seen, to be liked and to be shared - before they lose their sparks. And thus we can conclude, by watching these images through the visualizations created for this projects, that the images on Twitter posted during the World Cup were, above all, short-lived devices fighting for visibility.

4 Final Considerations

Our studies with large quantities of images still have a long way to go. With the current research, Labic has tried to bring question as well as contribute with new and already existing image visualizations methods. In a time where these data are easily produced and shared on every moment, initiatives that try to record these MOVEMENTS are even more needed, so narrative paths and rhythms can be discussed, despite the quick and uncertain life of the artifacts under consideration. Through Colors of the World Cup, we can observe the traces of images published in a network and try to determine its behavior, and by way of its vestiges comprehend its movements, apparitions and reapparations, and longevity after all.

However, there are still obstacles that tend to discourage this kind of research. The processing power needed for the conceptualization and execution of projects of this magnitude is very large: many files are used, and the large volume of images can lead machines to have a very low performance. Astable and high speed internet connection is also an important issue to insure that the image crawling can proceed without interruption.

Another constant problem faced in this kind of study is “bots”, robot-profiles that automatically share content and almost always in a massive way. This kind of automatic process can make specific network’s threads or subjects seem to have a high relevance, figuring between the most shared ones, though really their organic circulation is low: they are not shared by the network’s "common" users. For this reason it is crucial to have a watchful eye for bots, so they don’t taint the analysis.

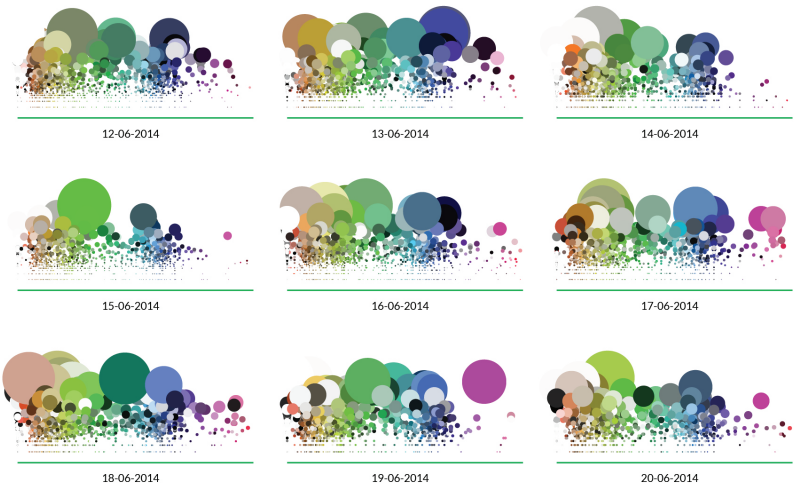


Fig. 1: Chromatic calendar: days 12-20 of June. Best viewed through the site.

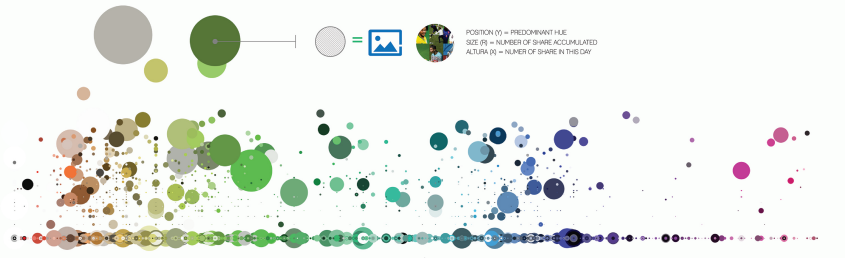


Fig. 2: Chromatic timeline. Best viewed through the site.

Colors of the World Cup provides an experimental method for data extraction in real time and introducing new image visualizations possibilities, offering a contribution to the cyberculture and image study fields and presenting novel ways to analyze image data.

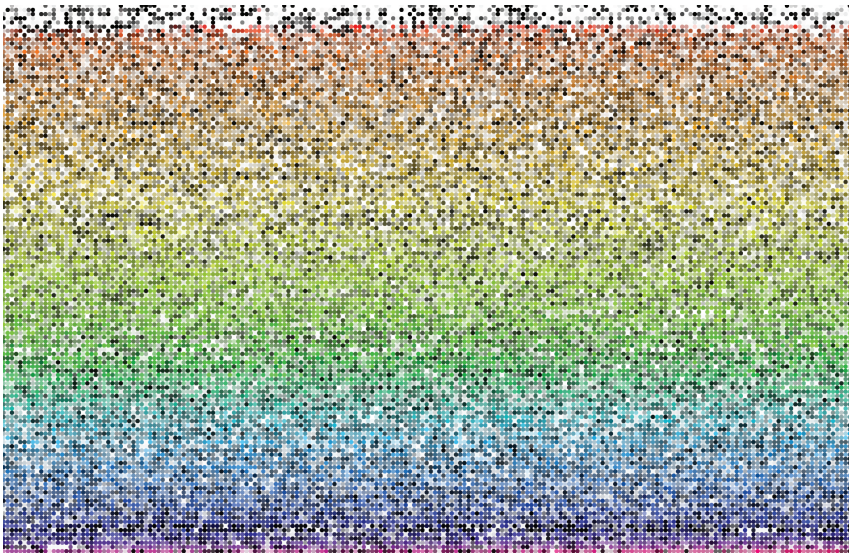


Fig. 3: Chromatic Mosaic. Best viewed through the site.

Fig. 4: Images Mosaic. Best viewed through the site.

Fig. 5: Most shared image of June 22. Author (s) unknown (s).

Fig. 6: Image composition with Neymar player, was the most shared in the period of the World Cup. Author (s) unknown (s).

Fig. 7: David Luiz (player) photo being consoled was the second most shared image in the day after the brazilian defeat to Germany. Author (s) unknown (s).

Bibliography

[1] Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009, April). Mapping the world's photos. In Proceedings of the 18th international conference on World wide web (pp. 761-770). ACM. Lee, S., Maisonneuve, N., Crandall, D., Efros, A. A., & Sivic, J. Linking Past to Present: Discovering Style in Two Centuries of Architecture. In *IEEE International Conference on Computational Photography*.

[2] Haacke, V. , Ribeiro, A. et. al. Do "Não Vai Ter Copa" para a "Copa dos Memes": Uma análise das imagens memes mais compartilhadas durante a Copa do Mundo FIFA 2014. Brazil, *Foz do Iguaçu: Intercom*, 2014.

[3] Hochman, N., & Manovich, L. (2013). Zooming into an Instagram City: Reading the local through social media. *First Monday*, 18(7).

[4] Manovich, L. (2012). How to compare one million images. *Understanding Digital Humanities*, 249-278.

Why Larry is different than Narry: A linguistic study of shipping communities in the One Direction fandom

Grace A. Ruiter

Calvin College
Grand Rapids, Michigan
graceruiter@gmail.com

Abstract. This study uses corpus linguistics, sentiment analysis, and discourse analysis to test the hypothesis that the Larry shippers—fans of the relationship between Louis and Harry from One Direction—behave differently than other shipping communities within the One Direction fandom. The study not only finds support for its hypothesis, but also reveals that the Larry shippers, who are suppressed for their support of a homosexual relationship, have themselves become oppressors through their treatment of another shipping community: the Elounor shippers. The Larry shippers dominate Twitter discussions about Louis Tomlinson and Eleanor Calder's relationship with attacks on the ship's validity. In their domination, the Larry shippers arguably disempower the official One Direction narrative about Elounor and replace it with their own interpretation of the relationship, an incredible feat for a fan community—one that would not be possible without social media avenues like Twitter

Keywords: fandom culture, corpus linguistics, sentiment analysis, discourse analysis

1 Introduction

Social media has given the general public an unprecedented ability to respond to important events in politics, entertainment, and business. With this ability comes the potential for groups to develop and perpetuate their own counter-narratives that reject mainstream ideas (Dahlberg, 2007). This democratization of the public sphere proves particularly significant for fandom communities. In fact, by enabling fans to respond to, and even reject "official" narratives, social media has fundamentally changed the relationship between cultural texts and their fans (Galuszka, 2014). Now more than ever, fans are not only consumers, but also creators, promoters, and critics in their own right.

One of the most popular ways fans interact with texts is through the practice of shipping, supporting a hypothetical or real romantic relationship between two characters or people. From Drarry (Draco Malfoy and Harry Potter), to JohnLock (Sherlock Holmes and John Watson), the social web plays host to a plethora of ships—many of which represent counter-discourses to the cultural texts that inspired them. However, because social media's emergence is recent, no previous linguistic research on fandom shipping communities and social media exists. Using the One Direction fandom as a case study, this study aims to set the groundwork for future scholarly work on the subject.

Few fandoms boast shipping communities as intense as the One Direction fandom's. And few fandoms can compare to the One Direction fandom in size and online activity. The most popular ship in that fandom, and the second most popular pairing for stories on the fan fiction site Ao3: Larry Stylinson (Harry Styles and Louis Tomlinson). Larry Stylinson is an unusual ship in that, officially, both Harry and Louis are heterosexual, yet almost everyone who ships Larry believes Harry and Louis are really in a secret relationship. The relationship is not public, shippers assert, because One Direction's management is forcing Harry and Louis to stay in the closet.

Larry shippers want their ship to be recognized as more than a fantasy. This study hypothesizes that, because of this desire, discussions about Larry follow different linguistic patterns than other discussions about ships within the One Direction fandom. In particular, the study anticipates that Larry shippers will employ more intentional rhetoric aimed at proving their ship is real than other shippers. The study tests this hypothesis by using corpus linguistics and discourse analysis tools to compare the content of Twitter discussions about Larry with discussions about five other ships within the One Direction fandom: Elounor (Louis Tomlinson and Eleanor Calder), Sophiam (Liam Payne and Sophia Smith), Zerrie (Zayn Malik and Perrie Edwards), Ziam (Zayn Malik and Liam Payne), and Narry (Harry Styles and Niall Horan).

2 Methodology

2.1 Data collection procedures

Materials. At the time of the study, three One Direction members were in public relationships with women: Zayn Malik, Liam Payne, and Louis Tomlinson. I included the ships centered on these relationships—Zerrie, Sophiam, and Elounor—in the study to represent the One Direction fandom's treatment of band members' publicly romantic relationships. The study compares these official heterosexual pairings with three ships centered on the publicly platonic relationships between band members. I selected Larry, Ziam, and Narry as the "bromance" ships for the study because they were the three most frequently discussed ships between One Direction band members on Twitter during December 2014.

Data collection procedure. To build corpora for the six ships in the study, I collected a random sample of 100 tweets—one tweet from each day between October 1, 2014 and January 8, 2015—for each of the ships included in the experiment. I used Topsy, a Twitter analytics tool, to search for the tweets that contained the names of the bromances and romances being studied. Topsy was set to display search results beginning with the tweets posted closest to the time of day at which I collected data. Collection took place between 12 and 9 p.m. Eastern Standard Time. I added the tweet at the top of the results list for each day to the corpus. I did not exclude or treat retweets differently in this process.

2.2 Data analysis procedures

Preparation for analysis. To build corpora for the six ships in the study, I collected a random sample of 100 tweets—one tweet from each day between October 1, 2014 and January 8, 2015—for each of the ships included in the experiment. I used Topsy, a Twitter analytics tool, to search for the tweets that contained the names of the bromances and romances being studied. Topsy was set to display search results beginning with the tweets posted closest to the time of day at which I collected data. Collection took place between 12 and 9 p.m. Eastern Standard Time. I added the tweet at the top of the results list for each day to the corpus. I did not exclude or treat retweets differently in this process.

Sentiment analysis. I classified each tweet in the six corpora as positive, neutral, or negative in its sentiments toward the ship or ships it discussed. I realize no language is truly neutral. However, I included a neutral category in my classification system because people in a fandom often have generally positive feelings toward ships within that fandom without actively shipping them. To get an accurate picture of attitudes toward each individual ship, I needed to divide those mildly positive sentiments from clearly positive attitudes. And in order to account for context, fandom-specific language, pictures, videos, GIFs, and emojis, I did all of the sentiment analysis without the use of computational systems. I made my classifications based on the definitions of positive, neutral, and negative provided below.

- **Positive sentiment**: Tweets that use clearly positive language, emoticons, or images to characterize the ship being discussed or those that count themselves fans of that ship.
- **Neutral sentiment**: Tweets that do not explicitly express positive or negative feelings about the discussed ship or those that ship it.
- **Negative sentiment**: Tweets that use negative language, images or rhetoric to describe a ship or those that ship that particular ship.

Tweet categories by sample

Organizational notes

The Elounor and Zerrie samples had a significant number of negative tweets. I broke the negative tweets for both of those samples down into individual types of criticism. However, none of the other four samples had more than 10 negative tweets. Because these samples had so few tweets with negative content, categorizing the tweets reveals little. Thus, although a study with a larger sample might be able to draw more conclusions about the rhetoric employed in negative tweets about those ships, for the purposes of this study, I chose to display the negative tweets in these samples as a single category.

The Elounor sample's general tweet data is also organized differently than the other samples. Because Larry shippers represented such a significant portion of that sample, the large volume of Larry shipper tweets skewed the data from other categories within the sample. To get an accurate representation of how each type of tweeter in the sample discussed Elounor, I broke down the Pro-Larry, Pro-Elounor, and Neutral tweet categories individually, in addition to the "By Category" results for the entire sample.

General tweet type categories:

- **Coming Out Discussion (Larry only)**: Discusses a hypothetical coming out for a shipped pair, implying belief in the ship's reality.
- **Creative Activity**: Displays creative interaction with a ship, but does not construct a fantasy based on the ship.
- **Defense of Ship**: Defends a ship and/or the act of shipping it.
- **Fangirl Behavior**: Uses hyperbolic, emotional language or images to discuss a ship.
- **Friendship (Narry only)**: Discusses a ship as a friendship.
- **General Fandom**: Mentions ship in the context of a whole fandom discussion.
- **Misc**: Does not fit in any established tweet categories.
- **Negative**: See "Negative Sentiment" definition in section 2.2b.
- **Obs/Moments (Evidence)**: Focuses on interactions between a pair being shipped.
- **Pro-Larry Elounor tweets (Elounor only)**: Discusses Elounor but is composed by a Larry shipper or Elounor skeptic.
- **"Relationship goals" tweets (Sophiam only)**: Describes a pairing as "relationship goals" for the tweeter.
- **Retweet/Fav if You Ship**: Asks fans to favorite or retweet as an expression of support for a pairing.
- **Sexual Content (Narry and Ziam only)**: Discusses ship in explicitly sexual way.
- **Ship as Fantasy**: Constructs a fantasy based on the ship.
- **Ship as Reality**: Explicitly discusses a ship as being "real," implying belief that the ship represents a real-life romance.
- **Spam**: Photos and GIFs of a person, a ship, or the whole band tweeted in rapid succession.
- **Tags**: Responds to questions posed in a tag that fandom members nominate each other to answer.

Negative tweet type categories:

- **Ad Hominem**: Attacks/insults those who ship a pairing.
- **Coming Out Discussion (Zerrie only)**: Discusses a hypothetical coming out for a shipped pair, implying belief in the ship's reality.
- **Comparisons (Elounor only)**: Makes unflattering comparisons between a pairing and other romantic relationships.
- **Disgust**: Expresses general distaste for a ship.
- **Dying Ship Narrative (Elounor only)**: Perpetuates narrative that Elounor is a dying ship, while the Larry fandom is growing.
- **Negative Fangirl Behavior (Zerrie only)**: Expresses negative sentiments toward a ship, but also displays fangirl behavior.
- **Predictions (Elounor only)**: Predicts future actions based on past events.
- **Proofs**: Subset of the Obs/Moments category in which a tweet uses events/interactions between a pairing as evidence for or against a ship's reality.
- **Skepticism (Zerrie only)**: Conveys doubts or suspicions about a ship.
- **Straw man**: Presents easily dismantled argument for Elounor, and uses images, GIFs, or text to dismiss the argument.

3 Results

3.1 Elounor Sample

Sentiment Analysis:

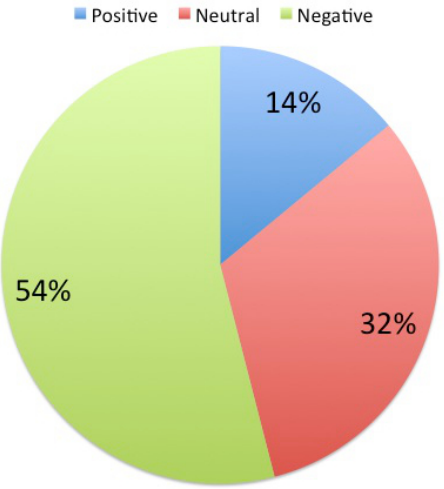


Fig. 1: Elounor Sentiment Analysis Results

Types of Fans Behind Tweets:

- Larry Shippers: 60%
- Neutral One Direction Fans: 26%
- Elounor Shippers: 14%

By Category:

- Pro-Larry Tweets: 62%
- Retweet/Fav If You Ship: 15%
- Fangirl Behavior: 9%
- Misc: 5%
- Obs/Moments: 4%
- Defense of Ship: 2%
- Creative Activity: 2%
- General Fandom: 1%

Pro-Larry Tweets (62% of sample):

- Neutral Sentiments About Elounor: 11%
- Negative Sentiments About Elounor: 89%

Negative (toward Elounor) Pro-Larry Tweets By Category:

- Straw Man: 24.1%
- Comparisons: 20.3%
- Dying Ship Narrative: 16.7%
- Proofs: 11.1%
- Ad Hominem: 11.1%
- Disgust: 9.3%
- Predictions: 7.4%

Elounor Shipper Tweets by Category (14% of sample):

- Fangirl Behavior: 69%
- Obs/Moments: 31%
- Creative Activity: 15%
- Defense of ship: 8%

Neutral Tweets by Category (24% of sample):

- Retweet/Fav If You Ship: 64%
- Misc: 20%
- Spam: 8%
- General Fandom: 4%
- Defense of Ship: 4%

3.2 Zerrie Sample

Sentiment Analysis:

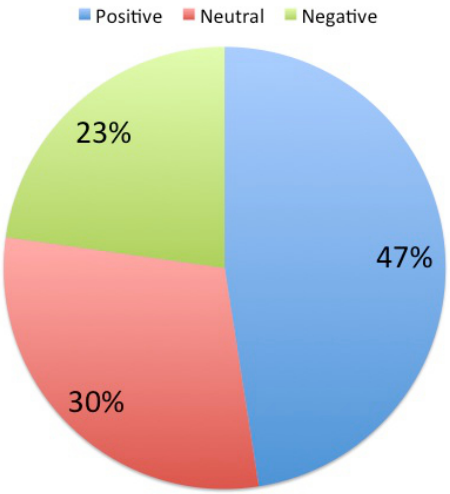


Fig. 2: Sophiam Sentiment Analysis Results

By Category:

- Fangirl Behavior: 40%
- Negative: 23%
- Proofs: 10%
- Defense of Ship: 8%
- Misc: 8%
- Retweet/Fav If You Ship: 7%
- Creative Activity: 6«
- General Fandom: 3%
- Ship As Reality: 2%
- Ship As Fantasy: 1%
- Spam: 1%
- Tags: 1%

Negative Tweets (toward Zerrie) By Category (23%):

- Skepticism/disbelief: 35%
- Disgust: 35%
- Ad Hominem Attacks: 26%
- Negative Fangirl Behavior: 26%
- Straw Man: 13%
- Proofs: 13%
- Coming Out Discussion: 4%

3.3 Sophiam Sample

Sentiment Analysis:

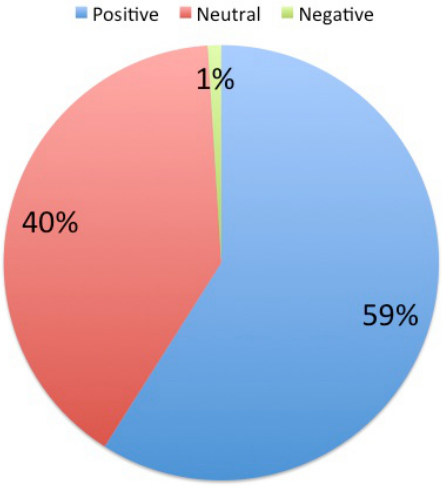


Fig. 3: Zerrie Sentiment Analysis Results

By Category:

- Fangirl Behavior: 47%
- "Relationship goals" tweets: 15%
- Retweet/Fav If You Ship: 14%
- Misc: 12%
- General Fandom: 9%
- Defense of Ship: 3%
- Reality of Ship: 2%
- Negative: 1%
- Tags: 1%
- Ship as Fantasy: 1%

3.4 Larry Sample

Sentiment Analysis:

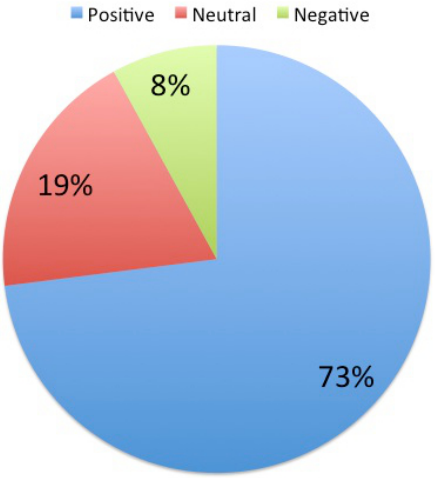


Fig. 4: Larry Sentiment Analysis Results

By Category:

- Fangirl Behavior: 43%
- Ship as Reality: 17%
- Obs/Moments: 14%
- Negative: 8%
- Spam 8%
- Coming Out Discussion: 7%
- Misc: 6%
- Defense of Ship: 5%
- Ship as Fantasy: 5%
- Tags: 2%
- General Fandom: 1%

3.5 Narry Sample

Sentiment Analysis:

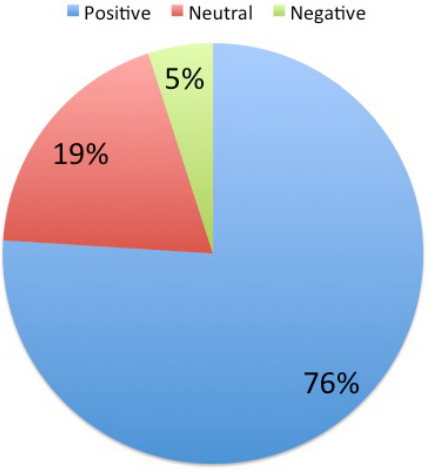


Fig. 5: Narry Sentiment Analysis Results

By Category:

- Fangirl Behavior: 67%
- Obs/Moments: 31%
- Friendship Tweets: 9%
- Misc: 7%
- Negative: 5%
- Fantasy: 5%
- Spam: 4%
- Sexual Content: 4%
- Retweet/Fav If You Ship: 3%
- Tags: 2%
- General Fandom: 1%

3.6 Ziam Sample

Sentiment Analysis:

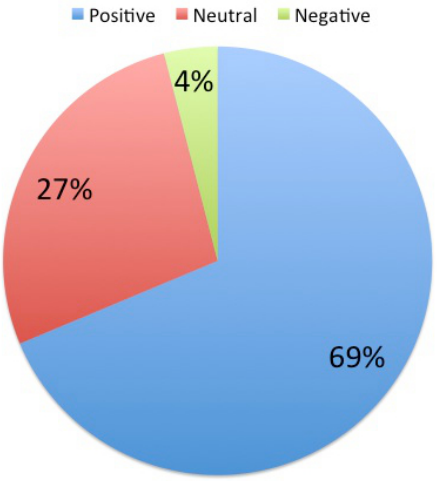


Fig. 6: Ziam Sentiment Analysis Results

By Category:

- Fangirl Behavior: 59%
- Spam: 2%
- Obs/Moments: 18%
- Fantasy: 9%
- Misc: 8%
- Retweet/Fav If You Ship: 6%
- General Fandom: 6%
- Negative: 4%
- Ziam as Reality: 4%
- Tags: 2%
- Sexual Content: 2%
- Creative Activity: 2%

4 Discussion

The experiment confirmed the hypothesis that discussions about Larry demonstrate different linguistic behaviors than discussions about other ships within the One Direction fandom on Twitter. However, it was the Elounor sample, not the Larry sample, which provided the strongest evidence of a difference between Larry and other shipping communities in the One Direction fandom.

Most Larry shippers know what it is like to be bullied for their beliefs. Shipping a homosexual relationship between two members of a high-profile band is not an act that

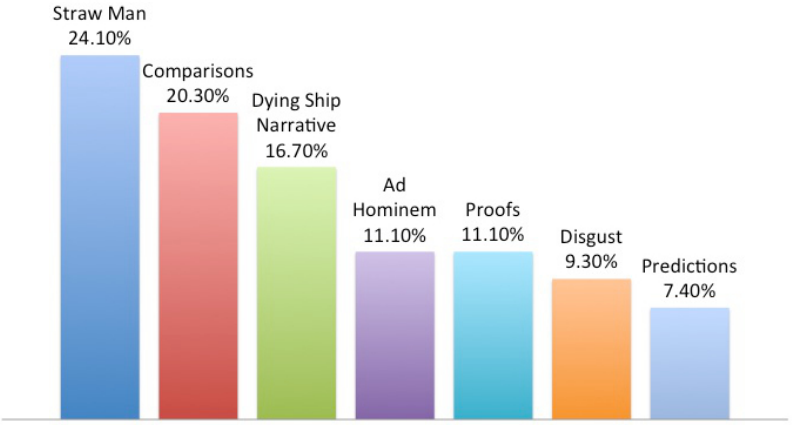


Fig. 7: Negative Tweets in the Elounor sample by category.

conforms to the heteronormative mainstream. On the contrary, those who publicly identify as LGBTQ+ or express support for the LGBTQ+ community often experience significant marginalization and suppression. Additionally, public participation in a fandom on any level often exposes people to shaming and mockery by the mainstream (Bennett, 2010). Yet the Elounor sample reveals that Larry shippers have, upon becoming a dominant group within the One Direction fandom, themselves become oppressors.

54% of the tweets in the Elounor sample expressed clearly negative sentiments toward Elounor and those ship it. Nearly all of those negative tweets could be traced back to a Larry shipper. Larry shipper tweets accounted for 60% of the tweets collected for the Elounor sample. 89% of those tweets had clearly negative sentiments toward Elounor.

It has been theorized that the development of opposition narratives on the Internet has a positive impact, fostering a more democratic model of communication and breaking down powerful hegemonies (Dahlberg, 2007). However, the bullying behaviors Larry shippers display suggest that, if oppositional discourses develop enough power, they may fall into the same cycle of oppression as the discourses they seek to dismantle.

The Larry shippers on Twitter utilize classic rhetorical strategies to paint Elounor as the fantasy and Larry as the reality. In the process, they become bully-victims—defending their beliefs from suppression by becoming the suppressors:

In the most blatantly aggressive tweets, the Larry shippers launch ad hominem attacks on the character of Elounor shippers. For example, one Larry shipper tweeted, “elounor shippers parents: ’if all of your friends jumped off a bridge would you do it?’ elounor shippers: ‘if they said so.”’ This tweet paints the Elounor shippers as naïve and gullible. And by extension, it portrays Larry shippers as enlightened for seeing Elounor as fake. The tweet offers no evidence for Larry or against Elounor to back up these claims.

Although not all of the pro-Larry tweets are as explicitly aggressive as the ad hominem attacks, when broken down, they, too, depend more on rhetorical strategies than concrete evidence and arguments to make their point. Of the seven strategies identified in the sample,

straw man arguments proved most popular, accounting for 24% of the pro-Larry tweets. With this strategy, the Larry shippers took advantage of the lack of vocal Elounor defenders in the Twitter community by replacing their arguments with basic statements about Elounor, statements that could easily be torn apart. For example, one shipper tweeted, "Elounor is real," but paired the tweet with a screenshot of Harry's New Year's tweet: "It's 2015." This implied that, by now, it should be obvious to all that Larry, not Elounor, is the real relationship. As with the ad hominem attacks, this strategy did not address real arguments for or against Elounor.

The second most popular approach saw shippers compare affectionate images of Sophiam or Zerrie to images of Louis and Elounor looking bored to suggest that Elounor does not look "loved-up" enough to be a real couple. The evidence many of these tweets provide amounts to a manipulated interpretation of isolated images. With hundreds of pap pictures to choose from, it is easy for Larry shippers to find candid pictures in which Elounor look unhappy, as well as images in which Sophiam and Zerrie appear very much in love. These moments hardly constitute reality. However, in employing rhetorical strategies like this one, it is possible that the Larry shippers have successfully influenced the perception of Larry and Elounor within the fandom.

Conclusions about how many people actively ship a given pairing are beyond the scope of this study. That said, just 14% of the fans in the Elounor sample expressed clear appreciation of the relationship. Every other sample contained significantly more positive sentiments toward the ship discussed than negative. In fact, four of the six samples had less than ten negative tweets. If a majority of One Direction fans still believe Elounor to be a real relationship, it appears they lack the passion and numbers to represent it on social media. By contrast, the Larry fandom has amassed enough social media influence to account for more tweets about Elounor—a ship none of the Larry shippers actively support—than Elounor shippers themselves. Thus, whether the Larry shippers directly discouraged fans from expressing appreciation for Elounor or not, they have successfully disempowered the official One Direction narrative, an incredible feat for a fan community—and one that would not be possible without social media avenues like Twitter. This is an excellent example of the way social media has fundamentally changed the fan-artist relationship.

The Larry sample itself adds further credence to the hypothesis that Larry shippers display different linguistic behaviors than other shipping communities. Notably, 17% of the tweets in the Larry sample contained a variation of the phrase "Larry is real." While several other samples did discuss the reality of a given ship, such discussion occurred with far less frequency. The Larry sample also featured discussions of a hypothetical "coming out," implying belief that Louis and Harry were really in love and would eventually announce it to the world. Although one tweet in the Zerrie sample discussed Ziam "coming out," there were no other mentions of ships besides Larry coming out in any of the samples.

Taken together with the behavior displayed in the Elounor sample, these differences observed in the Larry sample support the hypothesis that Larry shippers engage in linguistic behaviors distinct from other shipping communities within the fandom and consistent with a desire for recognition of their ship as reality.

In general, all of the samples but the Elounor sample contained more positive sentiments than negative. The samples also shared high levels of typical fangirl behavior—in some of the samples, more than half of the tweets display fangirl behavior or language. There were, however, a few differences between the discussions about Band Member + Band Member pairings and romances. The romance samples contained more neutral sentiments, suggesting that the enthusiasm for romances in the One Direction fandom is not as high as that for bromances, even if fans accept the romantic relationships. Additionally, the tweets in the Narry, Ziam, and Larry samples were more likely to feature reactions to interactions between the band members being shipped, or observations that connected to the ship. Perhaps because the ships centered on actual romances are less popular and more private, this form of fan interaction did not occur as often in any of those samples.

The tweets in the Narry sample distinguish that community even further from romance-focused ships. 9% of the tweets in the Narry sample specifically define Narry as a friendship, rather than a romance. While the Ziam sample contains more romantically focused content than the Narry sample, many of the Ziam tweets also portray Ziam as fantasy. The Ziam sample has more tweets focused on fiction constructed around the ship than any of the other samples. Furthermore, a number of tweets in the Ziam sample display tolerance of multiple ships. Unlike Larry shippers, who only ship Harry and Louis together, Ziam shippers appear comfortable with ships between Zayn and Liam and females, as well as other band mates. The fact that Ziam shippers acknowledge and accept Liam and Zayn's romantic relationships suggests that most Ziam shippers do not see Ziam as a real-life romance. This represents an important difference in mindset between the Larry shippers and the Ziam and Narry shippers. Larry shippers do see their ship as reality, and this study affirms the hypothesis that their linguistic behavior reflects that viewpoint.

5 Conclusions

This study supports the hypothesis Larry discussions on Twitter possess statistically significant linguistic differences from other shipping discussions. Larry shippers assert their ship's reality at much higher rates than do the other groups in the study, and they are the only group to frequently attack another shipping community in their tweets. The study finds that, by attacking Elounor, Larry shippers have used their numbers and influence within the One Direction fandom to successfully take control of the Elounor narrative on Twitter, dominating discussions about Elounor with rhetoric that undermines the ship's validity. This is a monumental achievement for a fan community. However, the study also reveals that, in the process, the Larry shippers, who represent a belief system that is suppressed by the heteronormative mainstream, have themselves become oppressors. Within the One Direction fandom, more research on fandom behavior over different periods of time is needed. This study only looked at Twitter behavior between September and December 2014. Incidentally, Louis Tomlinson and Eleanor Calder broke up in March 2015, which could drastically change the way in which Larry shippers discuss Elounor. Further research might examine how discussions of Elounor have shifted in light of the break up, as well as how conversations may have evolved in the years leading up to the time span on which this study focuses. Additional

research could also delve into the linguistic behavior and sentiments within the fandom and outside of it following Zayn Malik's departure from the band in March 2015.

Scholars interested in expanding the study beyond the One Direction fandom may pursue linguistic study of another fandom or people-group on social media to see if the behaviors are consistent with the findings of this study.

Bibliography

[1] Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from `http://www.laurenceanthony.net/`

[2] Bennett, C. (2010). Flaming the Fans: Shame and the Aesthetics of Queer Fandom in Todd Haynes's Velvet Goldmine. *Cinema Journal*, 49 (2), 17-39.

[3] Dahlberg, L. (2007). The Internet, deliberative democracy, and power: Radicalizing the public sphere. *International Journal of Media & Cultural Politics*, 3 (1), 47-64.

[4] Galuszka, P. (2014). New Economy of Fandom. *Popular Music and Society*, 38 (1), 25-43.

[5] Twitter Search, Monitoring, & Analytics. (2014). Retrieved from `http://topsy.com/`

A friend, not a phone call: Brands on Twitter and the future of customer service

Josephine Bromley

University of Leeds
Leeds, UK

Abstract. It is clear that social media has had a huge impact on the relationship that consumers and companies have. Consumers expect more and, with the widespread availability and adoption of social media technology, are able to demand more from companies. Evaluating the language that companies use on social media can therefore indicate what strategies companies employ in order to manage the changing relationship between consumers and customers, in the face of the new expectations placed on them. This research suggests that while companies tend to employ language reflective of an individual, human identity in order to establish the relationship of a friend, others continue to maintain a corporate tone. The efficacy of both methods is discussed, suggesting further research into consumer opinion of companies on social media.

Keywords: corpus linguistics, customer service, customer experience, brand identity

1 Introduction

Social media has become a ubiquitous tool for connectivity, with two out of the top three most used websites globally being social media sites (Alexa Internet, 2015), along with almost one quarter of the world's population registered as users on social media sites in 2014 (eMarketer, 2013). Indeed, it is certainly not unreasonable to claim that "[t]oday, everything is about social media" (Kaplan & Haenlein, 2010 p 67) and just a couple of poignant examples, such as social media's role in the Arab Spring (Khondker, 2011; Eltantawy & Wiest, 2011; Howard & Hussain, 2011), as well as in the success of the 2008 Obama presidential campaign (Cogburn & Espinoza-Vasquez, 2011; Harfoush, 2009), indicate the pivotal influence social media can have. This study focuses on the effect social media has had on the identity of brands that use this type of media n order to communicate with and connect to consumers, and ultimately what this means for the future of customer service. Section 1 deals with an overview of the change in identity for brands on social media, as well as detailing the data studied and methodology used. Sections 2 and 3 involve analysis of the data and arguments for particular brand identities. Finally, Section 4 and 5 conclude the paper with a discussion of consumers' reaction to communicating with brands on social media and ow this relates to the future of customer service.

1.1 The impact of social media

The arrival and wide spread uptake of social media has had a huge impact on the relationship between companies and consumers. Social media has empowered consumers not only in having their voice heard, as noted by Deighton and Kornfeld (2009, p 4) "digital innovations of the last decade have made it effortless, indeed second nature, for audiences to talk back", but also to even become promoters and distributers of their own offers, with Hennig-Thurau et al (2010, p 311) citing examples such as video content on YouTube and products on eBay. The same authors (2010, p 311) continue to write that this development serves to threaten established business models, with a salient example being the crisis traditional print newspapers face regarding profitability in the digital age (Harper, 2010, p 1). This increase in consumer agency may be seen to be affecting a change in the powerful identity of businesses in a company-consumer relationship that is correlational; as consumers become more empowered, traditional businesses become less authoritative, "the key thing [. . .] is how social networking shifts power, and in the business context that means shifting power from producer to consumer" (Breakenridge, 2008, p 132). This has consequently caused the identity of the company to shift from a one-way transmitter of products and services to consumers, to a more equal participant in a dialogue, as Holt and Perren (2009, p 8) note, "the 'broadcasting model' of one-way communication from a powerful central agency to dispersed and mutually disconnected passive consumers, is giving way at an unprecedented pace to a 'broadband model"', where, "consumers are linked in social networks, and productive energy comes from anywhere in the system".

As a result of this, businesses are arguably forced to establish an identity on these "social networks" that will be accepted by the newly empowered consumer, if they hope to communicate successfully with consumers and build a relationship with them. This is particularly true for consumer-facing, branded businesses, as they rely on an understanding of their customers in order to be able to market and sell their products effectively. Arockiaraj and Baranidharan (2013, p 473) echo this idea when they write "[t]o ensure their brands survival, the brand owners have to seek an ever-greater understanding of what the consumer wants and to develop a relationship between them and the brand". Indeed, to disregard the importance of the new platform can even be said to be detrimental to the future success of consumer-facing businesses, as Qualman (2009, p 13) notes "while we may only be at the beginning of this revolution [. . .] it is imperative for social media to be an integral part of a company's overall strategy". The impendence of this "revolution" may be therefore seen as a key incentive for brands to engage with social media, as it is an investment in marketing the brand to an ever growing proportion of consumers.

In order to communicate effectively with the empowered consumers on social media, brands are required to define a more personable identity. To this, Fournier and Avery (2011, p 193) note that "as more branding activity moves to the web, marketers are confronted with the stark realisation that social media was made for people, not for brands". Indeed, in order for customers on social media to engage with brands, Ramsay (2010, p 258) writes that humanised communication is an important factor, "[f]ailure to humanise contact [. . .] can result in very low or no engagement around a campaign". Malhotra, Malhotra-Kubowicz,

and See (2012) also advise brands communicating with consumers to "[h]umanise the brand" and "[i]nject emotions", as they argue that consumers "like messages that paint the brand as a living object and express human emotions". This humanisation of communication can be seen to lead to a consumer feeling at 'one' with the brand (Engeseth, 2005, as cited by Yan, 2011, p 690). Yan (2011, p 691) further quotes Engeseth when he writes, "Engeseth's theory follows that the separate nature of many brand relationships – the 'them' and 'us' – is now obsolete", resulting in the distinction between brand and consumer, in terms of identity, becoming "blurred" (Yan, 2011, p 691), further reflecting the notion of the equalising nature of social media on the power between company and consumer. In this way, the identity of brands on social media may be seen to be that of a personable entity that communicates as an equal to the consumer. In order to investigate the ideas surrounding this humanised identity of brands on social media, the research questions this study hopes to explore are:

- What identity do brands have on social media?, and
- How is this identity established?

In order to answer these questions, I will analyse the language that brands use on social media in order to establish their identity on the platform. In the first instance, creating a successful brand identity is a precise process and company marketing is therefore a carefully considered process (Armstrong & Kotler, 2015, p 33), which includes of course the marketing language used. This investment in developing a brand identity has a clear economic incentive, as Erdogmus and Cicek (2012, p 1354) explain succinctly:

> "As brands gain exclusive, positive, and prominent meaning in the minds of a large number of consumers, they become irresistible and irreplaceable, and win the loyalty of the consumers. Brand loyalty, in return, brings sales revenues, market share, profitability to the firms, and helps them grow or at least maintain themselves in the marketplace."

True, the meaning that brands have for consumers can take precedence over the product, as may be evidenced in the discussion of the data used for study, with three separate bread brands appearing in the top five best-selling fast moving consumer goods (FMCG) brands of 2014. Finally, and central to the reason for analysing the language brands use, as brands' communication with consumers on social media is a key research focus for this study, then it is appropriate to study what these brands are actually posting on the platform.

1.2 The data

Social media cannot be defined by a particular genre, but instead should be viewed as encompassing outlets that are numerous and varied. For example, Mangold and Faulds (2009, p 358) cite examples of social media outlets such as social networking sites (Facebook, MySpace), company-sponsored websites and blogs (Apple.com, P&G's Vocalpoint) and collaborative websites (Wikipedia). The social media outlet Twitter is defined as "a microblogging service" where "users tweet about any topic within the 140-character limit and follow others to receive their tweets" (Kwak, Lee, Park & Moon, 2010, p 591). Established in 2006 (Carlson, 2011),

Twitter now has an estimated 288 million users, with an average 10 million new users joining the platform every year's quarter since 2010 (Statista, 2015a), highlighting again the massive growth of social media.

Twitter was chosen as the social media outlet of study for this paper because it lends itself well to the research focus of company-consumer relationships and the identity of brands on social media. Indeed, Twitter is a key platform "for dialogic communication between companies and consumers" (Lin & Peña, 2013 p 17), which entails that companies must be able to establish a clear brand identity in order to engage customers. The data comes from the Twitter accounts of the top 10 best-selling FMCG brands in the United Kingdom (UK) from 2014 according to Brand Footprint 2014, an online data source which "reveals how consumers around the world today are buying FMCG brands" (Kantar Worldpanel, 2014a). FMCG brands were focussed on because their target audience is everyday consumers, such as those that use social media sites like Twitter. Indeed, the site is also one of the most popular forms of social media in the UK as of 2014 (Statista, 2015b). The data was collected over a time period of the first two years of each brand's Twitter use, as the brands started using the platform at different times. The full breakdown of the data is displayed in Table 1, with the order of the brands representing their ranking in Brand Footprint 2014 (Kantar Worldpanel, 2014b).

Table 1: Breakdown of data in corpus.

Company	**Time Period**	**Tokens**
Warburtons	10/2011 – 09/2013	78,245
Heinz	04/2013 – 11/2014	28,791
McVitie's	10/2010 – 09/2012	31,425
Hovis	12/2010 – 11/2012	9,859
Kingsmill	12/2011 – 11/2013	11,928
Birds Eye	05/2013 – 11/2014	43,922
Müller	07/2011 – 06/2013	2,840
Walkers	01/2009 – 12/2010	199,550
Coca Cola	12/2011 – 11/2013	82,157
Cadbury's	07/2010 – 06/2011	240,460
Total Corpus Size		**729,177**

1.3 Difficulties with the data

As is always the case when working with real instance of language use, the data is not entirely 'clean'. For instance, not all brands use Twitter to the same extent, with two of the brands' Twitter use starting only in 2013, meaning that there was not two years' worth of data available for collection for every brand (Table 1), as the data was collected in December 2014. Further, not all brands used Twitter to the same extent, again resulting in an inequality

between the sub-corpora. Nevertheless, despite these imperfections, the corpus still provided adequate data for analysis, and methodological practises, such as comparing the data in terms of percentages relative to the sub-corpora, have arguably served to remedy the issue of differences in sub-corpora size.

1.4 Methodology

Biber, Connor and Upton (2007, p 2) write that "[c]orpus linguistic studies are generally considered to be a type of discourse analysis because they describe the use of linguistic forms in context". A corpus linguistic approach was chosen as the method for the study of the discourse of brands in a social media context because it allows for generalisations to be made about real instances of language use, due to the representative nature of the text sample (Biber, Connor & Upton, 2007, p 3). Baker (2006, p 13) also concludes that corpora are useful in providing better evidence for "an underlying hegemonic discourse" than relying on intuition about a single text sample alone. Furthermore, corpora allow for researcher bias to be reduced because it becomes less easy to be selective about certain texts, "we at least are able to place a number of restrictions on our cognitive biases" (Baker, 2006, p 12).

McEnery and Hardie (2012, p 1) also note "corpus linguistics is a heterogeneous field" and a mixture of methodologies within corpus linguistics is often utilised in order to more convincingly address the research question/s. For this study, a mixture of both quantitative and qualitative analysis have been used, for example by analysing the frequency of a certain word across sub-corpora, as well as investigating specific uses of the word in concordance lists in order to gain a fuller picture of how the word is used. This follows the benefits of triangulation, in that "more than one method should be used in the validation process to ensure that the variance reflected that of the trait and not of the method" (Cambell & Fiske, 1959, as cited by Jick, 1979, p 602). In order to explore the data, the corpus analysis software Wordsmith (Scott, 2012) was utilised, as it allows for this mixture of both qualitative and quantitative analysis.

2 Customer service

In the 2015 report for The Definitive Guide to Social Customer Service (Conversocial, 2015, p 5), it is written that "social media is changing the entire business of customer service, posing great challenges and presenting new opportunities for brands". This may be seen because it is causing a shift, as outlined in Section 1, from one-way communication from a powerful company, to more equalised, humanised communication. The report notes this communication as being one of the five pillars of the concept of "social first", placing social media at the centre of customer engagement (Conversocial, 2015, p 13), with the pillar in question being to "[c]onnect with customers on a deeply personal and emotional level to build relationships and trust" (Conversocial, 2015, p 13), thereby reflecting this notion of a more humanised identity.

2.1 Customers and consumers

Twitter may be seen as a key platform on which companies can provide customer service to consumers also using this form of social media. Indeed, even on the Twitter website itself, it is suggested that providing customer service is a major reason for businesses to be on Twitter (Twitter Basics, 2015). Thomases (2010, p 280) writes that Twitter can be very conductive to delivering certain aspects of customer service, and that reaching out on the platform to help customers with issues and problems is a focal point for many brands' Twitter strategy. The reasons for using Twitter in this way comprise benefits for both company and consumer. For example, the real-time conversational nature of the platform means that customer issues can be handled immediately, which is favourable for consumers, while the transparent nature of companies' responses allows companies to demonstrate their expertise and commitment to customer service in a visible and impactful way (Postman, 2009, p 54). An initial search of the words 'customer*' and the synonymous 'consumer*', with the asterix indicating that this search included plural forms of the words, in the entire corpus revealed that the words occurrences amount to 0.04% and 0.06% respectively. Figure 1 displays the usage of both words in percentages according to each brand, with the percentage on top referring to occurrences of 'consumer*' and the percentage beneath to 'customer*'. Although some brands may be seen to use these words more than others, with Heinz's use of 'consumer' being particularly salient, when these words are used in general across the entire corpus they can be seen to collocate almost exclusively in reference to entities related to customer services (Table 2), which would seem to reflect this method of using Twitter to rectify customer issues. Furthermore, the word 'sorry' appears as the 80th most frequent word in the corpus, which may again indicate the common utilisation of Twitter as a method of providing customer service, as sorry suggests that the companies are apologising for something. This ranking increases to 64th position in the keyword list for the corpus, indicating that its usage is particularly salient. Baker (2006, p 125) writes that a keyword list give a measure of saliency, rather than simply frequency, as it measures the frequency of a particular word's occurrence in general language use against that in the corpus being studied. This serves to provide a greater indication of notable usage of the word being examined. The keyword list for the corpus of this study was created by comparing it with a sample of the British National Corpus, "a 100 million word collection of samples of written and spoken language [. . .], designed to represent a wide cross-section of British English" (The British National Corpus, 2007). Figure 2 displays the distribution of 'sorry' in each sub-corpus. As may be seen in this distribution, as well as indicated in the previous Figure 1 and Table 2, Heinz can be seen to be a main contributor to the semantic field of customer service present in the corpus, along with the brands Kingsmill and Birds Eye, as suggested in Figure 2.

Although these brands may be seen to be responsible for the majority of language related to customer service, most brands in the corpus may be seen to provide some form of customer service. Table 3 displays qualitative examples of this provision of customer service by the brands. Regarding what this means for the identity of brand, this arguably results in a somewhat discrepant effect on a more humanised identity. Although there are logically beneficial reasons for using Twitter in this way, as outlined above, this may be

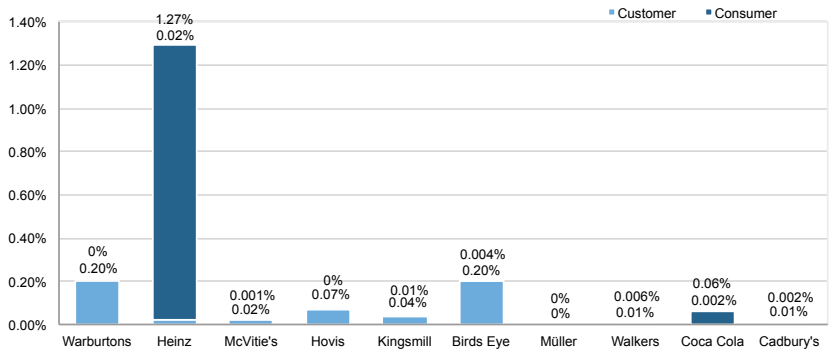


Fig. 1: Percentage of occurrence of 'customer*' and 'consumer*' according to each sub-corpus.

Table 2: R1 collocates of 'customer*' and 'consumer*' in the entire corpus.

Word	**R1 Collocate**	**Percentage**
customer	care	40.5%
	relation[1]	2%
	service	53.6%
		Total 96.1%
consumer	care	2%
	response@cokecce	7%
	information centre	3%
	relations	0.6%
	care@hjheinz.com	81%
	services@unitedbiscuits.com	0.5%
	@walkers.co.uk	3%
		Total 97.1%

seen to construct a situation where brands uphold a depersonalised, corporate identity when communicating with consumers. In support of this notion of the impersonal identity of corporations, Toth and Heath (2009, p 166) write that "the study of communication in the corporate or organisational context makes especially challenging questions about [...] the degree of 'personalness"', noting interestingly as well that the corresponding terms for 'Inc.' in French and Spanish are société anonyme and sociedad anónima, "[t]he corporation is then the anonymous society" (Toth & Heath, 2009, p 166). Sheth, Sisodia and Wolfe (2007, p 229) further note that traditional corporate company values do not particularly identify with "the human species", quoting to this William Vanderbilt, the 19th century American railroad magnate, "[t]he public be damned. I am working for my stockholders". Although not as overtly disregarding of consumers as William Vanderbilt, some brands in the corpus may be seen to indeed communicate with consumers in an impersonal manner. For instance,

focussing on the concordance list for 'sorry' in the Heinz sub-corpus, it is revealed that the phrase:

> "Sorry to hear, please call us on 08005285757 or email your telephone number to consumercare@uk.hjheinz so we can call you."

is repeated 276 times, accounting therefore for 58% of all occurrences of 'sorry' in the Heinz sub-corpus. It is clear that this tweet from Heinz is a canned response, noted as being disheartening to customers as the customer raising the issue feels they are not considered on a personal basis (Performance Research Associates, 2012, p 130). Kingsmill may also be seen to use such a practise of a mechanical, rather than human response, as 75% of the instances of 'sorry' occur in the same repeated phrase (Figure 3). The repetitive nature of these responses may be seen to reflect McGuinn's (2009) point that the foundation of customer service is mechanistic in nature, noting that this is a key weakness of customer service because it fails to treat the consumer as an individual, "[t]he customer is viewed in indirect terms, with a focus on collective customers rather than individual customers" (McGuinn, 2009, p 61). In this way, customer service interactions would seem to affect the communication with consumers negatively, with regards to a humanised brand identity, as the language can become impersonal and indifferent to the customer's emotional satisfaction. Indeed, Meyer and Schwager (2007, p 1) also note that customer service tends to be dispassionate, "a paucity of the personal touch [is] evidence of indifference to what should be a company's first concern: the quality of customers' experiences".

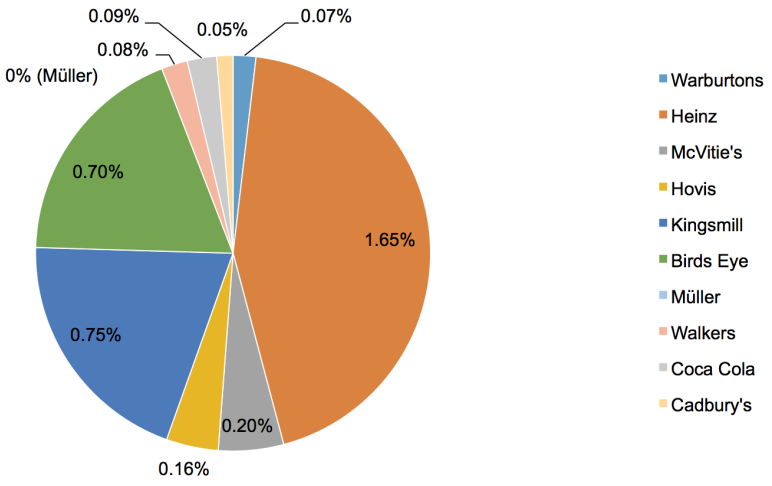


Fig. 2: Distribution of 'sorry' among sub-corpora.

Table 3: Examples of brands providing customer service.

Brand	Example
Warburtons	"Hi! That doesn't sound right. Could you contact customer care team? 0800 243 684 or talk2us@warburtons.co.uk"
Heinz	"We're sorry to hear about this. Please contact us on 0800 5285757. We'd appreciate the chance to talk. Consumer Care"
McVitie's	"Sorry to hear that. Please contact customer services on vipclubsupport@coinks.com and they'll sort it out for you."
Hovis	"That does not look good. Do not fear, we will forward it to our customer services dept."
Kingsmill	"Hi there, we've checked with customer services and we replied on the 4th Dec. We will send it out again as you didn't get it."
Birds Eye	"We'll pass your email address onto our Consumer Care team and they'll get in contact with you. Thank you."
Walkers	"Hmmm... That's strange. In that case, it's best if you contact our customer care team... Ŝam"
Coca Cola	"Hi Simon, that does sound unusual! You can talk to our Consumer Information Centre about it on 0800 22 77 11 (press option 3)"
Cadbury's	"how very disappointing! There's a customer service contact in the pack that should be able to help you..."

2.2 Grammar of apologies

The way in which Heinz apologises arguably has the effect of mitigating the degree of fault the company assumes, which may be seen to echo this notion of apathetic communication of corporate companies. For example, the list of 3-word clusters, which Biber, Conrad and Cortez (2004, p 372) explain are "multi-word prefabricated expressions" that function similarly to single word units, that are generated in the concordance list of 'sorry' in the Heinz sub-corpus reveals that the phrase 'sorry to hear' is the most frequent cluster to collocate with 'sorry' (Figure 4), occurring 83% of the time. Arguably, this distances the company from the responsibility for the customer's issue, as the syntax of the phrase would seem to be noting a state of sorrow, rather than explicitly apologising, as observed by Battistella (2014, p 58), "[s]aying 'I'm sorry' is different from saying 'I apologise'. The former reports on an internal state of the speaker but does not literally perform an apology". The author (2014, p 58) continues that the grammar of 'sorry' allows for an infinitive to be used, and if this is an infinitive verb of perception, as in 'sorry to hear', then this is interpreted as expressing empathy rather than an apology. In this way, the point that Heinz is 'sorry to hear' about the customer's problem, but is not 'sorry for' the problem serves to remove Heinz's responsibility from the issue. Indeed, there are only five instances of the verb 'apologise' in the entire corpus (Example 1), and none of these occur in the sub-corpora of Heinz, Birds Eye or Kingsmill, the brands that would seem to lean most towards using Twitter as a platform to provide customer service. The only brands that do use this verb are Walkers and McVitie's, with only McVitie's explicitly using the verb in response to a customer complaint, which may

be evidenced by the fact that they are replacing “dented tins”. In fact, there is barely any explicit ownership of mistakes the company has made which have caused the customer to be dissatisfied in any collocations of ‘sorry’ in the Heinz sub-corpus. This may be found in terms of transitivity, a grammatical system which “construes the world of experience into a set of process types” (Halliday, 2004, p 170). There are just two occurrences of material processes in the Heinz sub-corpus which collocate with ‘sorry’ (Table 4), with the vast majority being relational processes, with the difference between the two process types being that material clauses are “processes of doing & happening” and relational clauses are “processes of being & having” (Halliday, 2004). This gives the impression that Heinz has not done anything and has made no mistake, but rather that they simply are ‘sorry’ that this has happened. The instances of ‘sorry’ in the Kingsmill and Birds Eye sub-corpora were also examined in order to check the number of material process types, and it was found that there was only one occurrence of a material process for Birds Eye (Table 4), which may be seen to be in collocation with a humorous use of the word ‘sorry’ rather than an apologetic use, and no occurrences for Kingsmill. Ultimately, this may be viewed as a dispassionate response from the company, as it does not provide consumers with a true sense that they have received an apology. In this way, the language use of these brands in particular may be seen to display incongruent behaviour regarding establishing a personable, humanised identity on social media.

hi Laura We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
Hi Dani We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
@FivewayFilms We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
Hi Lucy We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
Hi Chris We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
@ee61re: We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
@carolineyes10: We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
@KingsmillCare: We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks
Hi Aimee We are sorry to hear this. Please send your details to careline@lovekingsmill.co.uk and we will reimburse you. Thanks

Fig. 3: Example of canned response from Kingsmill sub-corpus.

N	Cluster	Freq.	Length
1	**SORRY TO HEAR**	**396**	**3**
2	TO HEAR PLEASE	278	3
3	HEAR PLEASE CALL	273	3
4	PLEASE CALL US	250	3
5	TO HEAR ABOUT	57	3

Fig. 4: Top five most frequent 3-word clusters to collocate with ‘sorry’ in Heinz sub-corpus.

Example 1. All instances of 'apologise' from entire corpus.

"@walkers_crisps 13 Apr 2010 @kohsamui14 That's very true! I do apologise! ;) Ŝam"
"@walkers_crisps 14 Apr 2010 @RosieAllOver I do apologise! ;o) I think you should buy some new flavours and end the temptation! It's only right! Ŝam"
"@McVities 13 Aug 2012 @DJMOG1 Oh no, can't apologise enough for that! Will get a new one (make that 2!) out to you asap, we have your details"
"@walkers_crisps 13 Apr 2010 @MosherAngel No need to apologise. Thanks for the feedback :) Ŝam"
"@McVities 13 Aug 2012 @DJMOG1 No, we apologise for all the probs so far, you seem to have been extra unlucky! 2 new dent free tins will be with you assap.[2]"

This lack of humanisation in the identity of the brands discussed above may be seen as a failure in utilising the full capacity of Twitter to provide a more humanised customer service. As mentioned at the beginning of this section, social media is changing customer service, from "anonymous, one-to-one channels" toward "public, one-to-many channels that are [...] attached to real identity" (Conversocial, 2015, p 4). Indeed, in The Future of Customer Service report by trendwatching (2014, p 15) it is predicted that "webcam enabled face-to-face interaction" with a customer service representative will be a growing practise within companies' customer service remit, thereby providing a definite human interaction. The report writes that this trend echoes the imperative for the customer to feel "[l]istened to, valued and cared for" (trendwatching, 2014, p 4). This acknowledgement of the importance of customers' emotions is also noted by Frow and Payne (2007, p 91), when writing about creating an "'outstanding' or 'perfect' customer experience", in that companies "will need to consider the creation of customers' experiences from both rational and emotional perspectives". Note the authors' use of 'experience' rather than 'service'. Although this may be seen as a pedantic semantic point, by using the word 'experience', this arguably highlights the distinction between the impersonal communication manifested in the language use of the brands above and a more personable, humanised interaction. This type of interaction and the effect this has on the identity of the brands in the corpus are explored in Section 3.

Table 4: Material processes that collocate with 'sorry' in Heinz and Birds Eye sub-corpora.

Heinz	sorry	we no longer produce any of the puddings
		we don't make Big Soup pots anymore
Birds Eye		we never carry cash – just waffles

[2] spelling error made by brand, not by the author of this study.

3 Customer experience

Moving on from the conceivably depersonalised identities that are established by the brands discussed in Section 2, other brands in the corpus may be seen to be active in establishing an identity more in line with the humanised identities of brands on social media in order to engage consumers, with this engagement also arguably more proactive in relation to the reactive nature of brands when managing customer service issues.

Initial evidence for this engagement may be seen in the form of asking questions to involve the customer in a dialogue, which is a traditional advertising technique of "simulating the processes, structure and dynamism of everyday conversation" (Frank, 1989, p 255). For example, Figure 5 displays the percentage of questions which include the second person pronoun in all its forms, including 'you', 'your', 'yours' and 'yourself', out of all occurrences of questions in each sub-corpus, with the percentages above each column indicating the total percentage occurrence of questions across the sub-corpora. By using the second person pronoun, the brands may be seen to be using direct address in order to engage the consumer, "generic forms of 'you'/'your' draw audiences into the discourse" (Pollach, 2005, p 296). Moreover, questions including reference to the second person serve to establish a dialogue with the referee, as it is inherent in proposing a question that an answer is expected, as Thornborrow and Wareing observe (1998, p 192) "asking questions is a powerful way of addressing readers, since questions do not usually occur without a potential answer". As Figure 5 denotes, the majority of questions asked by all brands in the corpus contain direct address, with 100% of questions posed by Heinz referring to the second person. This may be seen to be as a result of Heinz's preoccupation of using Twitter mainly as a platform to provide customer service, with the majority of tweets asking customers to send their details (e.g. "Sorry to hear that. Help us report this to our team. Can you send us a DM with the codes from the can & your name & address?"). This prevalence of questions including direct address can be seen to reflect the brands' attempts to proactively engage the consumer in a dialogue "questions [. . .] have a direct appeal in bringing the second person into a kind of dialogue with the writer" (Webber, 1994, as cited by Thompson, 1998, p 138), in order to build an interactive relationship with them.

3.1 Content of tweets

After engaging the customer into conversation, the content of the brands' tweets also reveals that it is geared towards consumer engagement with the brand. For example, in the word list for entire corpus 'win' appears as the 72nd most frequent word. In order to examine all forms of the words, the distribution of the 'win' lexeme between the sub-corpora was quantified (Figure 6). Again, the keyword list for the corpus was referred to, which further evidenced the saliency of the word 'win', as the word appears in 55th position. The word is also the most frequent lexical item in terms of the language the companies have chosen to use, rather than lexical items that are inherent in the site themselves. For instance, the words 'view' and 'conversation' appear much more frequently than 'win' in 11th and 12th position respectively, but this is due to a functionality of Twitter, which allows users to view a series

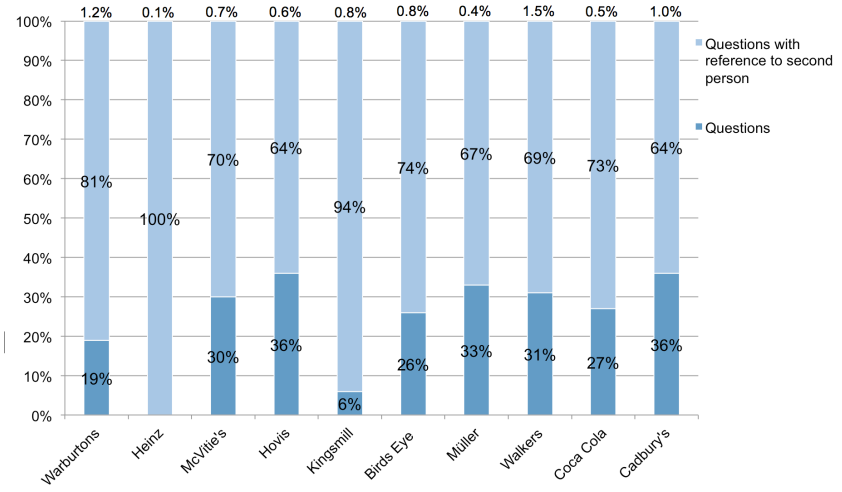


Fig. 5: Percentage of questions and questions with reference to second person among sub-corpora.

of tweets between particular users. Mangold and Fauld (2009, p 362) suggest that examples of customer-engaging promotions include contests, online games and online competitions. Following this, the salient occurrence of 'win' across the entire corpus indicates the way in which companies attempt to engage with consumers on social media, by inviting them to enter competitions to win prizes. For example, the most common prize that can be won may be seen to be 'tickets', as it is the second most frequent noun collocate, after 'chance', to occur with 'win' (Figure 7).

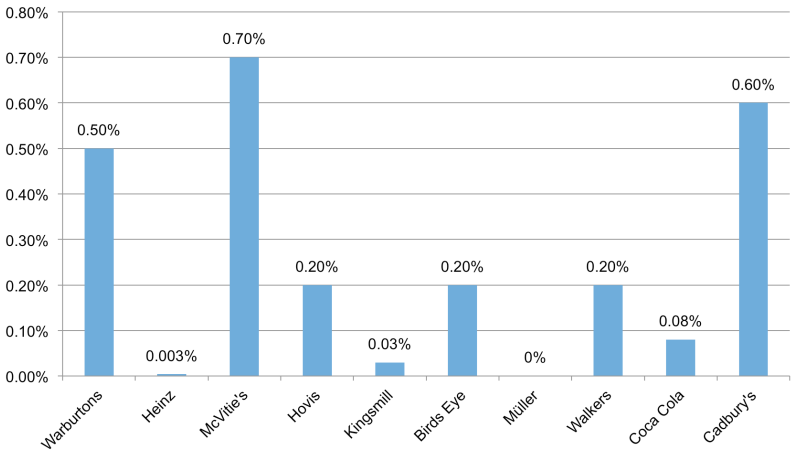


Fig. 6: Percentage distribution of 'win' lexeme.

Along with competitions and prizes, games played simply for the enjoyment of winning can be seen to be an engagement strategy employed. A notable instance of this was instigated by Cadbury's, when they invited users to take part in a Tweetoff every Thursday. Users could play along by simply tweeting "#Spots" or "#Stripes", dependent on which side, either Spots or Stripes, they chose to support. The winning team would be that which achieved the greatest number of tweeted hashtags by a certain time. Although prizes could be won by randomly selected members of the winning team, a further incentive for playing was arguably the possibility of simply winning the game, by collecting points for the chosen team. Indeed, Lee (2011, p 118) notes that "playfulness is a core activity in many new media contexts", and the specific mode of Twitter can be seen to lend itself well to this type of customer engagement of mass participation in a game. Because of the hashtag function on Twitter, which indicates a topic of discussion (Twitter, 2014), users are able to instantly join a particular topic and therefore, as in the case of the Tweetoff, participate in a game. Indeed, Cadbury's may be categorised as being one of the more playful brands in the corpus, with the brand using the lexeme 'win' the second most frequently, behind McVitie's, as well as accounting for 59% of all overall occurrences of the word 'win', contributing heavily therefore to the word's notable position in the word list for the entire corpus. However, it is worth to bear in mind that the weekly Tweetoff was part of the wider Spots v Stripes campaign, launched in 2011 to coincide with the London 2012 Olympic Games (Bainbridge, 2011), as Cadbury's was an official sponsor for the Olympic Games (BBC News, 2008). Therefore the brand's heavy use of the word 'win' and organisation of the mass Tweetoff game may be said to be due to a vested interest in promoting London 2012, due to the obligations of sponsorship, rather than purely establishing themselves as having a playful identity. More evidence for this may be seen in the high frequency of the collocate '2012' (Figure 7), which may be seen to be in reference to London 2012. This is not to say, of course, that the decision to sponsor London 2012 was not made without considering the impact on the identity of the Cadbury's brand, but rather that this sponsorship may be said to have allowed for a playful identity on Twitter, rather than Cadbury's establishing a playful identity on the platform in the first instance.

N	Word	With	R	Texts	Total	Total L	Total R	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	WIN	win	0.000	1	1,676	11	11	8	2	1			1,654			1	2	8
2	TO	win	0.000	1	1,097	970	127	34	18	16	5	897			39	34	33	21
3	A	win	0.000	1	841	255	586	10	31	194	14	6		524	15	8	14	25
4	YOU	win	0.000	1	349	304	45	12	49	63	145	35		11	3	11	13	7
5	THE	win	0.000	1	302	156	146	42	32	73	6	3		66	12	24	21	23
6	OF	win	0.000	1	285	58	227	18	25	13	1	1			52	86	75	14
7	0	win	0.000	1	283	0	283							79	8	61	23	112
8	FOR	win	0.000	1	257	206	51	17	174	13	2			11	9	9	11	11
9	2012	win	0.000	1	238	88	150	22	6	23	2	35		4	112	10	15	9
10	CHANCE	win	0.000	1	236	235	1	1	5	4	225				1			
11	HTTP	win	0.000	1	233	87	146	7	80					15	12	14	67	38
12	TICKETS	win	0.000	1	227	8	219	4	3		1			46	36	102	27	8

Fig. 7: Screen shot of top 11 collocates of 'win'.

Nevertheless, as Figure 6 suggests, other brands in the corpus would seem to employ an element of playfulness in establishing their identity on Twitter. Take for example, McVitie's, not a sponsor of London 2012, and the most prolific user of the lexeme 'win' in the corpus (Figure 6). The prizes available to be won by McVitie's if a user takes part in "#FreebieFriday", the name of the hashtag associated with competitions on the McVitie's profile, are varied and may be seen to be more often than not unrelated to McVitie's products. For example, Table 5 displays the most frequent noun collocates of 'win' before explicit mention of the type of products McVitie's sells, 'biscuits'. The right-hand collocation position of nouns such as 'iPhone', 'cinema' and 'tickets' indicate that these are the objects that are available to be won, due to the transitive nature of the verb to win, in that it may take an object (Hurford, 1994, p 242), with the qualitative examples from the McVitie's sub-corpus underneath Table 5 suggesting evidence for this indication (Example 2). The point that prizes which bear little relation to the product that McVitie's sells, "the UK's favourite biscuits and cakes" (United Biscuits, 2014), may suggest that the activity of playing and engaging with consumers with this way takes precedence over explicit promotion of the product.

Table 5: Most frequent noun collocations of 'win' before 'biscuits' collocation in McVitie's sub-corpus.

Collocate	**Frequency of Left-hand Collocation**	**Frequency of Right-hand Collocation**	**Total Frequency**
Chance	30.5%	0%	30.5%
Draw	19.7%	1.2%	20.9%
iPhone	0%	8.2%	8.2%
Cinema	0%	6.4%	6.4%
Tickets	0%	6.4%	6.4%
Time	1.9%	3.8%	5.7%
Day	3.2%	2.5%	5.7%
Voucher	0%	5.0%	5.0%
Box	0%	5.0%	5.0%
Mug	0.6%	3.9%	4.5%
Pair	0%	4.5%	4.5%
Biscuits	0%	3.9%	3.9%

Example 2. Qualitative examples of right-hand collocates of 'win'.

"Wow, this must be your lucky week! You should enter our draw to win 2 cinema tickets w/popcorn & drinks"

"Today's draw is for a Kindle!!! Get over to http://mcvitiesvipclub.co.uk to enter + you'll go in the Dec 25 draw to win an iPhone 4S!"

This may also be reflective of a recommended Twitter strategy of companies on the platform in order to engage consumers. Sierra (2012, as cited by O'Reilly & Milstein 2012,

p 223) says that "[w]ith few exceptions, the worst mistake a 'business blog' can make is to blog about the business", while Kaplan and Haenlein (2011, p 110) note that "the first rule of micro-blogging is to focus on messages that are relevant for the target group". Indeed, this strategy may be seen in the content of the brands' tweets. For instance, Birds Eye uses the hashtag '#foodjoke' after writing a food-related joke (Example 3), which may be seen as a way of engaging customers, by tweeting about things related to the brand's products, rather than the actual products themselves. The use of humour may furthermore be seen as an engagement strategy, "[p]eople like to laugh, funny things are appreciated" (Malhotra, Malhotra-Kubowicz & See, 2012), and simultaneously works towards establishing an identity for the brand that is more human. To this, Brown (2012, p 41) writes that a sense of humour allows customers to discover that the company is staffed by "real people", as opposed to "mechanoids without a soul", while Burkhardt (2010, p 12) notes that, in regards to communication with "a personal touch" on social media, "[b]eing human involves humour".

Arguably, this technique of tweeting about things unrelated to the product in a playful sense serves to reflect a more humanised identity, as the emphasis on sales and profitability is reduced because the consumer is not being overtly and directly sold the product. Instead, the consumer is presented with an invitation to engage with the brand under more light-hearted circumstances, such as by playing a game or entering a competition, ultimately allowing the brand to establish an identity which is more personable and even more likeable.

Therefore, consumers interacting with these brands on Twitter are arguably provided a proactive experience, which is in contrast to the reactive, depersonalised service interaction as explored in Section 2.

Example 3. All instances of '#foodjoke' in Birds Eye sub-corpus.

#FoodJoke: What's the fastest vegetable? A runner bean!
"Why did the tomato blush? Because it saw the salad dressing! #foodjoke"
"Why do the french like to eat snails? Because they don't like fast food. #foodjoke"
"Waiter, will my pizza be long? No sir, it will be round! #foodjoke"
"What is black, white, green and bumpy? A pickle wearing a tuxedo! #foodjoke"

Example 4. All instances of first person singular pronoun use in Müller sub-corpus.

"@FreshMilkDaily 18 Mar 2013 @ironfarmer86 John – dm me with the contact details and I'll put farm services in touch." "@FreshMilkDaily 10 Mar 2013 @Rob_Hitch a little confused. Did you message @freshmilkdaily? I've passed your tweet to farm services – they'll be in touch." "@Fresh Milk Daily 14 Nov 2011 @chriswalkland chris – sitting on your email now. We'll take username and password off I think. . . " "@FreshMilkDaily 6 Apr 2013 @ironfarmer86 sort with your contacts in the marketing team. I've passed your tweet to them."

3.2 Pronoun use

This personable identity may be further evidenced in the brands' use of pronouns. Waugh (2010, p 82) writes that "[i]n the identity literature there has been some attention given to personal pronouns like I, you, we, they as important indicators of identity". De Fina (1995, p 380) also references studies that have looked at "pronouns within social diexis", citing Levinson (1983), who writes "the social identities of participants" are systematically encoded in pronominal paradigms (De Fina, 1995, p 380). In this way, observing the pronoun use of brands in the corpus can help to determine the humanised identity of brands on Twitter.

In a word list of the entire corpus, the second person pronoun 'you' appears around 30% more frequently than the first person plural pronoun 'we' (Figure 8). This may again be seen as an established engagement strategy by using direct address, "[advertisement's] use of 'you' is part of a high-involvement strategy which attempts to win us over by very direct address" (Cook, 1992, p 157). However, what may strike as noteworthy about the pronoun use in the corpus is that around 1/7th of first and second person pronoun use refers to the first person singular, resulting in a greater occurrence of first person pronoun use overall. This may be said to subvert the expected use of pronouns that may be seen in traditional marketing language, which can be said to focus more on the consumer. Indeed, this is inferred by Cook (1992, p 156) when writing about pronouns in advertising, "most striking and most frequent [. . .] and most divergent from the uses of other discourse types, is the ubiquitous use of 'you"', with Kalmane (2012, p 87) also noting that in advertising "[t]he second person pronouns are used more than first person pronouns". In order to explore this atypical use of pronouns further, the brands' pronoun use was broken down into further detail (Table 6). The figures of first and second person pronoun use for each sub-corpus include use of pronouns in all forms, which includes personal pronouns (e.g. 'I'), reflexive pronouns (e.g. 'ourselves'), possessive pronouns functioning as adjectives (e.g. 'your') and possessive pronouns functioning as nouns (e.g. 'yours'). As displayed in the table, the majority of brands would seem to conform to the expected first and second pronoun usage in advertising discourse, with the majority of pronouns being in the second person. Yet, it is clear that some brands do not follow this convention, and it may be said that none have an overwhelming majority of second person pronoun usage that is suggested in the literature, "[a]dvertisements are peppered with you's and I's and we's and your's – especially the you's" (Sedivy & Carlson, 2011, p 165). This notable distribution of pronouns may therefore be interpreted as the brands' attempt at establishing a more human identity. In a similar study to this one, Kwon and Sung (2013) studied elements of brand anthropomorphism, including pronouns, and found that, just as may be viewed in Table 6, there was roughly a 50-50 split between first person pronouns and second person pronouns, "[a]pproximately 54% of the tweets contained personal pronouns: 736 with second-person and 650 with first-person pronouns" (Kwon and Sung, 2013, p 11). The researchers noted that by using pronouns in this way, "marketers try to imbue human personality into their brands" (Kwon & Sung, 2013, p 4). Thus, in the same way, by regularly using the first person pronoun, the brands can be seen to be consistently marking a humanised identity.

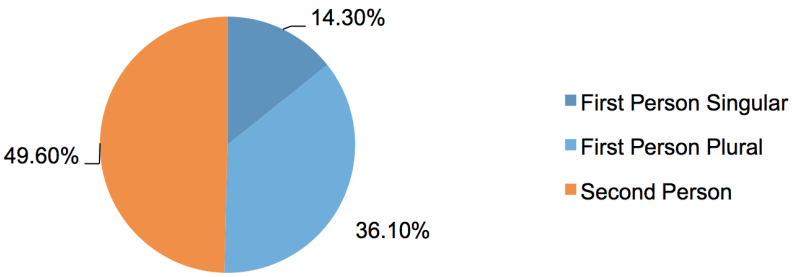


Fig. 8: Total pronoun use in entire corpus.

The notable occurrence of 'I' in the corpus may be seen to further this notion of a "human personality", according to Small, Geldart, Gutman, and Clarke-Scott (1998, p 298) "the use of first person pronouns [. . .] is an expression of one's personal identity". It may be worth to point out that all brands apart from Kingsmill use a mixture of both first person singular and first person plural (Table 6), which may be seen to be as a phenomena related to the discourse of Twitter. Boyd, Golder and Lotan (2010, p 5) write that regarding the conversational aspects of Twitter "[a]mbiguities abound [. . .] with the respect to pronoun usage in the content of messages". This may be due to differences in the conversational function of initial tweets in comparison to replies. Indeed, referring again to Kwon and Sung's study (2013), it was found that the first person was more common in tweeted replies, "more anthropomorphism variables occurred in replies than original tweets" (Kwon & Sung, 2013, p 14). True, Müller displays this kind of variability of pronouns in replying to tweets (Example 4). By writing, for example, "We'll take off your username and password I think", the writer of the tweet can be seen to be referring to both Müller as a whole company in the first person plural pronoun, while the first person singular pronoun refers to the writer themselves and their own uncertainty towards the statement about one of the company's practises, manifested in the use of epistemic modality in "think". In this way, these tweets from the Müller data with the inclusion of the first person singular pronoun gives the impression that a real person has written this response.

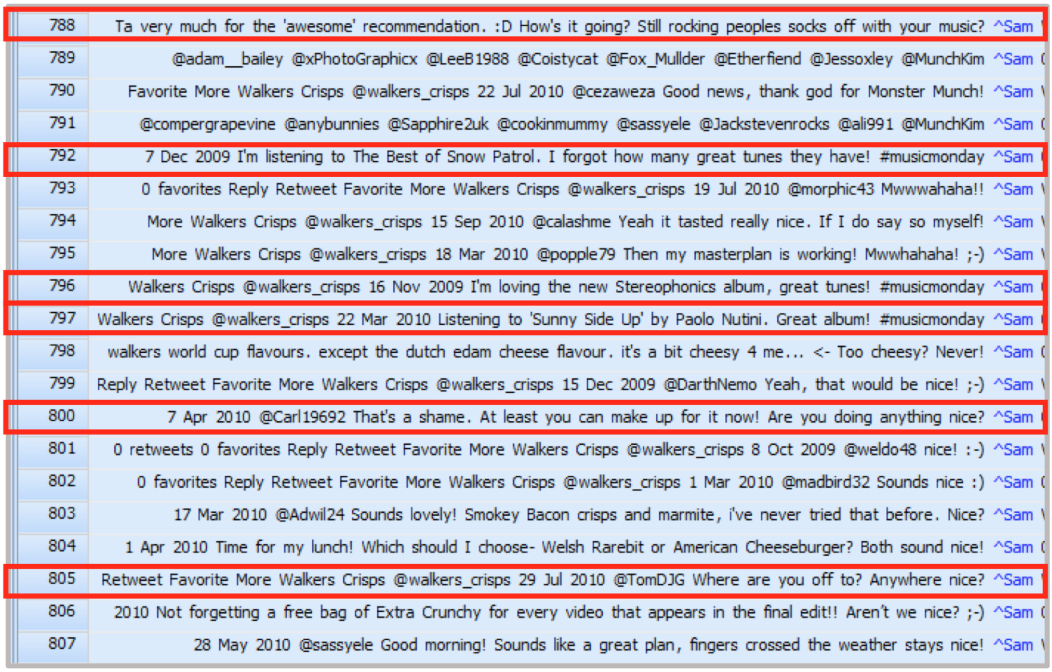
This perception of a person, and thus a human identity, may also be seen most prominently in the Walkers sub-corpus, which can be viewed in Table 6 to be an exceptional user of the first person singular, as it occurs close to four times as often as the first person plural. Closer analysis of the concordance line for 'I' in the Walkers sub-corpus reveals that this would seem to be accredited to a referee with a name: 'Sam' (Figure 9). Walkers appears to be the only brand in the corpus which employs this technique of including the tweet author's name. This may be viewed as a tactic intended to humanise the organisation and as a consequence the brand, as Pollach (2005, p 294) notes "[c]ompanies appeal to readers' emotions when they present the people behind the organisation" such as when "members are identified with names". Focussing on this qualitative example from the Walkers sub-corpus (Figure 9), some of these tweets by 'Sam' would seem to emulate the language that a real person could conceivably use. For example, Sam thanks another Twitter user for an "awesome

Table 6: First and second person pronoun use in each sub-corpus.

	Percentage of pronoun use out of total first and second person pronoun use for each sub-corpus		
	First Person		Second Person
Brand	Singular (e.g. 'I', 'my')	Plural (e.g. 'we', 'our')	(e.g. 'you', 'your')
Warbutons	4.9%	44.3%	50.8%
Heinz	1.9%	42.7%	55.4%
McVitie's	1.2%	41.1%	57.7%
Hovis	7.1%	50.1%	42.8%
Kingsmill	0.0%	39.0%	61.0%
Birds Eye	2.0%	49.5%	48.5%
Müller	6.2%	47.7%	46.2%
Walkers	39.6%	10.3%	50.1%
Coca Cola	1.1%	45.5%	53.4%
Cadbury's	12.0%	43.8%	44.2%

recommendation" and asks if they are still performing music (Line 788), while declaring her/his enjoyment in listening to "Snow Patrol" (Line 792), "Stereophonics" (Line 796) and "Paolo Nutini" (Line 797). Sam also asks other Twitter users if they are doing anything nice (Line 800) and if they are going anywhere nice (Line 805). These tweets from Sam bear no relation to Walkers products, and arguably read as genuine "turns" in a conversation, suggesting a conversational sequence (Hutchby & Wooffitt, 2008, p 13), or remarks indicative of the discourse of social media, where users "confess their personal thoughts and feelings" (Zappavigna, 2012, p 28). This production of seemingly authentic discourse of a real person may be seen to be an extension of Fairclough's (2001) notion of "synthetic personalization", which is explained as "a compensatory tendency to give the impression of treating each of the people 'handled' en masse as an individual" (2001, p 52). With this in mind, 'Sam' can indeed be seen to connect with people at an individual level, while the public nature of conversations on Twitter arguably helps to give the impression that Walkers treats all people "en masse" in this individual manner. Fairclough (2001, p 184) further notes that "[a] common dimension of synthetic personalization is simulated equalization", which again may be reflected in the questions 'Sam' asks other Twitter users, "Are you doing anything nice?" (Line 800), seemingly simulating a discourse between friends, and thus 'equals', who take an interest in one another's activities. Ultimately, the conception of 'Sam' as a human representative for Walkers can be seen to afford the brand truly a humanised identity.

In this way, along with the other brands discussed in this section, Walkers may be seen to provide consumers with a customer experience, where the brand is imbued with human characteristics and where the consumer is treated on an individual, personalised basis, where their emotional status is taken into account ("How's it going?" (Line 788)) . As opposed to the brands analysed in Section 2, these brands may be seen to be employing the capability



788	Ta very much for the 'awesome' recommendation. :D How's it going? Still rocking peoples socks off with your music? ^Sam
789	@adam__bailey @xPhotoGraphicx @LeeB1988 @Coistycat @Fox_Mullder @Etherfiend @Jessoxley @MunchKim ^Sam
790	Favorite More Walkers Crisps @walkers_crisps 22 Jul 2010 @cezaweza Good news, thank god for Monster Munch! ^Sam
791	@compergrapevine @anybunnies @Sapphire2uk @cookinmummy @sassyele @Jackstevenrocks @ali991 @MunchKim ^Sam
792	7 Dec 2009 I'm listening to The Best of Snow Patrol. I forgot how many great tunes they have! #musicmonday ^Sam
793	0 favorites Reply Retweet Favorite More Walkers Crisps @walkers_crisps 19 Jul 2010 @morphic43 Mwwwahaha!! ^Sam
794	More Walkers Crisps @walkers_crisps 15 Sep 2010 @calashme Yeah it tasted really nice. If I do say so myself! ^Sam
795	More Walkers Crisps @walkers_crisps 18 Mar 2010 @popple79 Then my masterplan is working! Mwwhahaha! ;-) ^Sam
796	Walkers Crisps @walkers_crisps 16 Nov 2009 I'm loving the new Stereophonics album, great tunes! #musicmonday ^Sam
797	Walkers Crisps @walkers_crisps 22 Mar 2010 Listening to 'Sunny Side Up' by Paolo Nutini. Great album! #musicmonday ^Sam
798	walkers world cup flavours. except the dutch edam cheese flavour. it's a bit cheesy 4 me... <- Too cheesy? Never! ^Sam
799	Reply Retweet Favorite More Walkers Crisps @walkers_crisps 15 Dec 2009 @DarthNemo Yeah, that would be nice! ;-) ^Sam
800	7 Apr 2010 @Carl19692 That's a shame. At least you can make up for it now! Are you doing anything nice? ^Sam
801	0 retweets 0 favorites Reply Retweet Favorite More Walkers Crisps @walkers_crisps 8 Oct 2009 @weldo48 nice! :-) ^Sam
802	0 favorites Reply Retweet Favorite More Walkers Crisps @walkers_crisps 1 Mar 2010 @madbird32 Sounds nice :) ^Sam
803	17 Mar 2010 @Adwil24 Sounds lovely! Smokey Bacon crisps and marmite, i've never tried that before. Nice? ^Sam
804	1 Apr 2010 Time for my lunch! Which should I choose- Welsh Rarebit or American Cheeseburger? Both sound nice! ^Sam
805	Retweet Favorite More Walkers Crisps @walkers_crisps 29 Jul 2010 @TomDJG Where are you off to? Anywhere nice? ^Sam
806	2010 Not forgetting a free bag of Extra Crunchy for every video that appears in the final edit!! Aren't we nice? ;-) ^Sam
807	28 May 2010 @sassyele Good morning! Sounds like a great plan, fingers crossed the weather stays nice! ^Sam

Fig. 9: Qualitative extract from the concordance list for 'Sam' in the Walkers sub-corpus.

of Twitter to engage consumers by appealing to them on an emotional level, creating a personable humanised experience and therefore reflecting the future of customer service.

4 Discussion

It must be emphasised that the structure of this study, which separates brands that maintain a more detached corporate identity from brands that establish a more humanised identity, was not done with the intention of making definite black-and-white distinctions between the identities of the brands. As has hopefully been gathered by the quantitative displays of data from the whole corpus, it is clear that all brands use a mixture of customer service language and language more in line with the discourse of customer experience. Due to this, it may be worth to consider what effect this could have on the brands' identities. For example, referring back to Section 2, brands such as Walkers, which arguably has a very humanised brand identity, may also be seen to provide customer service on Twitter. An instance indicating Walkers' provision of customer service can be seen in the following tweet:

> "@Reeb198Hi. Sorry 2hear ur unhappy with the french fries. If you'd like 2 request a refund, pls call customer services on 0800 274 777 Ŝam"

By merging these two identities, as shown in this example, of a business service with the friend-like persona of 'Sam', this may perhaps cause discordance in the identity of the brand. O'Reilly and Milstein (2012, p 221) suggest that "opening an account or several just for customer service" may be a way in which to clearly direct consumers to register their

problems effectively, while De Beule (2014) also notes that there are certain cases where it is beneficial to have a separate customer service account. Nevertheless, the author maintains that "[n]o matter what kind of questions customers are dealing with, they seek to personally engage with a company" (De Beule, 2014). In this way, De Beule echoes the notion behind the title of this study, A friend, not a phone call, in that customers respond better to a humanised, known identity than detached customer service communication with a person they do not know.

Yet, in contrast to this argument, the opposite may also be suggested in that customers have been found to be unresponsive to, or even dislike, this friendly identity. For instance, Andriuzzi (2015) found that participants in a qualitative interview study did not particularly like when the brand attempted to engage consumers by discussing other topics unrelated to the company's products or services, while Baird and Parasnis (2011, p 1) discovered that "most [consumers] do not engage with companies via social media in order to feel connected", rather that they mostly engage in order to receive discounts on products. This pragmatism may be seen to be in contrast to the emotional connection that consumers are thought to desire. This may be seen to be related to the notion of social distance which "specifies the extent to which individuals have a close or a more distant relationship" (Meyer, 2010, p 64). For this reason, because consumers do not feel a true closeness to a brand in exactly the same way as they would to a friend, as suggested in the use of synthetic in "synthetic personalization", then this may perhaps be the reason why consumers may reject or ignore this type of engagement strategy from companies. In this way, the depersonalised identity of brands discussed in Section 2 may perhaps be a more comfortable and appropriate interaction for consumers who feel this way.

Table 7: Ranking in Brand Footprint 2014 and no. of followers on Twitter as of May 2015.

Companies in Brand Footprint 2014 order	$\neq$	**Companies in order of number of followers on Twitter**	**No. of followers**
Warburtons		Cadbury's	253,000
Heinz		Coca Cola	106,000
McVitie's		Walkers	51,300
Hovis		McVitie's	32,000
Kingsmill		Warburtons	19,800
Birds Eye		Birds Eye	13,200
Müller		Hovis	3848
Walkers		Müller	1509
Coca Cola		Heinz	927
Cadbury's		Kingsmill	714

With regards to the kind of consumer this may be, it would be of use to consider characteristics of the target audience for particular brands. For example, Table 7 shows

that Hovis places on the lower end of the scale in terms of number of followers on Twitter, which may be seen to indicate how engaged customers are with the brand on the platform, as Castronovo and Huang (2012, p 120) write that engagement is signified by "the size of the community". However, the average Twitter user in 2012 was an American female aged between 15-25 (beevolve, 2012), in comparison to Hovis' target audience, British women aged between 25-64 with children (Newsworks, 2013). Thus, it may be suggested that there still remains a disparity between the type of consumer actually purchasing the product and the type of Twitter user that brands could potentially engage with. Nevertheless, as predicted in Section 1 of this study, it may be ill-advised of brands to believe this will always be the case and a presence on the platform may still be seen as a necessary investment.

5 Conclusion

This study attempted to define the identities of the top 10 best-selling FMCG brands of 2014 on Twitter and how these are established using certain, salient linguistic features. For example, it was argued that by using the first person singular pronoun, brands can be seen to establish a more humanised identity, whereas brands that mitigate responsibility in the transitivity structure of their apologies may be viewed as having a more impersonal identity. The former identity can be seen to reflect trends in the future of customer service, while the latter may arguably fail to engage customers on social media in the same way.

However, in order to explore these conclusions further, as indicated in the previous Discussion Section, there is much to be said for further research into how consumers truly perceive brands on Twitter and other forms of social media, as well as consumers' reactions to brands' engagement strategies. This could be explored by analysing the discourse between brands and consumers, in order to gather an indication of what engagement strategies from companies that consumers respond to the most, as well as how consumers feel about this kind of communication. A continuation of a corpus linguistics approach may also be suitable in order to provide representative evidence of company-consumer interaction on social media, combined with instances of qualitative analysis, to be able to add richer detail to the research evidence.

Bibliography

[1] Alexa Internet. (2015). The top 500 sites on the web. Retrieved from: http://www.alexa.com/topsites

[2] Andriuzzi, A. (2015). The tweeting brand: When conversation leads to humanization. *Twitter Mix Days*, 24 April, Lyon.

[3] Armstrong, G. & Kotler, P. (2015). *Marketing: An Introduction*. 12th ed. Essex: Pearson Education.

[4] Arockiaraj, G. & Baranidharan, K. (2013). Impact of Social Media on Brand Awareness for Fast Moving Consumer Goods. *International Journal of Logistics & Supply Chain Management Perspectives*. 2(4). 472-477. Retrieved from: http://search.proquest.com/openview/fe07cd28f66db95a88cd42aa83fe0b57/1?pq-origsite=gscholar

[5] Bainbridge, J. (2011). Spots v Stripes: Cadbury versus the critics. Retrieved from: http://www.marketingmagazine.co.uk/article/1063866/spots-v-stripes-cadbury-versus-critics

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

[7] Battistella, E.L. (2014). *Sorry about that: The language of public apology*. Oxford: Oxford University Press.

[8] BBC News. (2008). Cadbury becomes Olympic sponsor. 20 October. Retrieved from: http://news.bbc.co.uk/1/hi/england/london/7679635.stm

[9] beevolve. (2012). An Exhaustive Study of Twitter Users Across the World. Retrieved from: http://www.beevolve.com/twitter-statistics/

[10] Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*. 25(3). 371-405. Retrieved from: http://www.corpus4u.org/forum/upload/forum/2005110322472496.pdf

[11] Biber, D., Connor, U. & Upton, T.A. (2007). *Discourse on the Move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins Publishing Co.

[12] Boyd, D., Golder, S. & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: *System Sciences (HICSS), 2010 43rd Hawaii International Conference on System Sciences*, 5-8 January 2010, Hawaii. Hawaii: IEEE Computer Society. 1-10. doi: 10.1109/HICSS.2010.412

[13] Breakenridge, D. (2008). *PR 2.0: New Media, New Tools, New Audiences*. New Jersey: Pearson Education.

[14] British National Corpus, The, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

[15] Brown, E. (2012). *Working the Crowd: Social media marketing for business*. 2nd ed. Swindon: British Informatics Society.

[16] Burkhardt, A. (2010). Social media: A guide for college and university libraries. *College & Research Libraries News*. 71(1). 10-24. Retrieved from: http://crln.acrl.org/content/71/1/10.short

[17] Carlson, N. (2011). The Real History of Twitter. *Business Insider*. 13 April. Retrieved from: http://www.businessinsider.com/how-twitter-was-founded-2011-4?IR=T

[18] Castronovo, C. & Huang, L. (2012). Social Media in an Alternative Marketing Communication Model. *Journal of Marketing Development and Competitiveness*. 6(1). 117-134. Retrieved from: http://www.na-businesspress.com/jmdc/castronovoc_web6_1_.pdf

[19] Cogburn, D.L. & Espinoza-Vasquez, F.K. (2011). From Networked Nominee to Networked Nation: Examining the Impact of Web 2.0 and Social Media on Political Participation and Civic Engagement in the 2008 Obama Campaign. *Journal of Political Marketing*,10, 189-213. doi: 10.1080/15377857.2011.540224

[20] Cook, G. (1992). The Discourse of Advertising. London: Routledge. Conversocial. (2015). *The Definitive Guide to Social Customer Service*. Retrieved from: http://www.conversocial.com/social-customer-service/introduction#.VUjWrvlVhHw

[21] De Beule. (2014). Should a Company Create a Separate Customer Service Twitter Handle? January 14. *engagor*. Retrieved from: https://engagor.com/blog/should-a-company-dedicate-a-separate-customer-service-twitter-handle/

[22] De Fina, A. (1995). Pronominal Choice, Identity and Solidarity in Political Discourse. Text – *Interdisciplinary Journal for the Study of Discourse*. 15(3). 379-410. doi: 10.1515/text.1.1995.15.3.379

[23] Deighton, J.A. & Kornfeld, L. (2009). Interactivity's Unanticipated Consequences for Marketers and Marketing. *Journal of Interactive Marketing*. 23(1). 2-12. doi: 10.1016/j.intmar.2008.10.001

[24] Eltantawy, N. & Wiest, J.B. (2011). Social Media in the Egyptian Revolution: Reconsidering Resource Mobilization Theory. *International Journal of Communications*. 5. 1207-1224. Retrieved from: http://ijoc.org/index.php/ijoc/article/view/1242/597

[25] eMarketer. (2013). Social Networking Reaches Nearly One in Four Around the World. *eMarketer*. Retrieved from: http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976

[26] Erdogmus, I.E. & Cicek, M. (2012). The impact of social media marketing on brand loyalty. *Procedia: Social and Behavioural Science*. 58. 1353-1360. doi: 10.1016/j.sbspro.2012.09.1119

[27] Fairclough, N. (2001). *Language and Power*. 2nd ed. Abingdon: Pearson Education.

[28] Fournier, S. & Avery, J. (2011). The uninvited brand. *Business Horizons*. 54. 193-207. doi: 10.1016/j.bushor.2011.01.001

[29] Frank, J. (1989). On conversational involvement by mail: The use of questions in direct sales letters. *Text and Talk*. 9(2). 231-259. doi: 10.1515/text.1.1989.9.2.231

[30] Halliday, M.A.K. (2004). *An Introduction to Functional Grammar*. London: Arnold.

[31] Harfoush, R. (2009). *Yes We Did! An inside look at how social media built the Obama brand*. Berkeley: Pearson Education.

[32] Harper, R. A. (2010). The Social Media Revolution: Exploring the Impact on Journalism and News Media Organizations. *Student Pulse*. 2(3). Retrieved from: http://www.studentpulse.com/a?id=202

[33] Hennig-Thurau, T., Malthouse, E.C., Friege, C. Gensler., Lobschat, L., Rangaswamy, A. & Skiera, B. (2010). The Impact of New Media on Customer Relationships. *Journal of Service Research*. 13(3). 311-330. doi: 10.1177/1094670510375460

[34] Holt, J & Perren, A. (2009). *Media Industries: History, Theory and Method*. Oxford: Wiley-Blackwell.

[35] Howard, P.N. & Hussain, M.M. (2011). The Role of Digital Media. *Journal of Democracy*. 22(3). 35-48. Retrieved from: http://muse.jhu.edu/login?auth=0&type=summary&url=/journals/journal_of_democracy/v022/22.3.howard.html

[36] Hurford, J.R. (1994). *Grammar: A Student's Guide*. Cambridge: Cambridge University Press.

[37] Hutchby, I & Wooffitt, R. (2008). *Conversation Analysis*. Cambridge: Polity Press.

[38] Jick, T.D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*. 24(4). 602-611. doi: 10.2307/2392366

[39] Kalmane, R. (2012). *Advertising: Using Words as Tools for Selling*. 2nd ed. Riga: Lulu Enterprises.

[40] Kantar Worldpanel. (2014a). What is Brand Footprint?. Retrieved from: http://www.brandfootprint-ranking.com/what-brand-footprint/

[41] Kantar Worldpanel. (2014b). Explore The Data: UK/FMCG Ranking. Retrieved from: http://www.brandfootprint-ranking.com/report/ranking/fmcg/2013,2012/country/uk/

[42] Kaplan, A.M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*. 54. 59-68. doi: 10.1016/j.bushor.2009.09.003

[43] Kaplan, A.M. & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*. 54. 105-113. doi: 10.1016/j.bushor.2010.09.004

[44] Khondker, H.H. (2011). Role of New Media in the Arab Spring. *Globalizations*. 8(5). 675-679. Retrieved from: http://www.tandfonline.com/doi/citedby/10.1080/14747731.2011.621287

[45] Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a Social Network or a News Media?. In: *WWW 2010, 26-30 April 2010, Raleigh. North Carolina: International World Wide Web Conference Committee*. 591 – 600. Retrieved from: http://www.eecs.wsu.edu/~assefaw/CptS580-06/papers/2010-www-twitter.pdf

[46] Kwon, E.S. & Sung, Y. (2013). Follow Me! Global Marketers' Twitter Use. *Journal of Interactive Advertising*. 12(1). 4-16. Retrieved from: http://jiad.org/article149.html

[47] Lin, J. & Peña, J. (2013). Are You Following Me? A Content Analysis of TV Networks' Brand Communication on Twitter. *Journal of Interactive Advertising*. 12(1). 17-29. Retrieved from: http://jiad.org/article150.html

[48] Mangold, W.G. & Faulds, D.J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*. 52. 357-365. doi: 10.1016/j.bushor.2009.03.002

[49] Malhotra, A., Malhotra-Kubowicz, C. & See, A. (2012). How to Create Brand Engagement on Facebook. *MIT Sloan Management Review*. 18 December. Retrieved from: http://sloanreview.mit.edu/article/how-to-create-brand-engagement-on-facebook/

[50] McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practise*. Cambridge: Cambridge University Press. Retrieved from: http://www.dawsonera.com

[51] McGuinn, C. (2009). The Future of Customer Service. *Irish Marketing Review*. 20(1). 57-66. Retrieved from: http://connection.ebscohost.com/c/articles/44050938/future-customer-service

[52] Meyer, C & Schwager, A. (2007). Understanding Customer Experience. *Harvard Business Review*. 1-11. Retrieved from: https://hbr.org/2007/02/understanding-customer-experience

[53] Meyer, C.F. (2010). *Introducing English Linguistics*. Cambridge: Cambridge University Press.

[54] Newsworks. (2013). Hovis: Case Study. Retrieved from: http://www.newsworks.org.uk/case-studies/67824

[55] O'Reilly, T & Milstein, S. (2012). *The Twitter Book*. 2nd ed. Sebastopol: O'Reilly Media.

[56] Performance Research Associates. (2012). *Delivering Knock Your Socks Off Service*. New York: American Management Association.

[57] Pollach, I. (2005). Corporate self-presentation on the WWW: Strategies for enhancing usability, credibility and utility. *Corporate Communications*. 10(4). 285-301. doi: http://dx.doi.org/10.1108/13563280510630098

[58] Postman, J. (2009). *SocialCorp: Social Media Goes Corporate*. Berkeley: Pearson Education.

[59] Qualman, E. (2009). *Socialnomics: How Social Media Transforms the Way We Live and Do Business*. New Jersey: John Wiley & Sons. Retrieved from: https://www.dawsonera.com

[60] Ramsay, M. (2010). Social media etiquette: A guide and checklist to the benefits and perils of social marketing. *Database Marketing & Consumer Strategy Management*. 17(3). 257-261. doi: 10.1057/dbm.2010.24

[61] Scott, M. (2012). *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software.

[62] Sedivy, J. & Carlson, G. (2011). *Sold on Language: How Advertisers Talk to You and What This Says About You*. Chichester: Wiley-Blackwell.

[63] Sheth, J. N., Sisodia, S.R. & Wolfe, B.D. (2007). *Firms of Endearment: How World-Class Companies Profit from Passion and Purpose*. New Jersey: Wharton School Publishing.

[64] Small, J.A., Geldart, K., Gutman, G. & Clarke-Scott, M-A. (1998). The discourse of self in dementia. *Ageing and Society*. 18(3). 291-316. doi: : http://dx.doi.org/ (About DOI)

[65] Statista. (2015a). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2014 (in millions). *Statista: The Statistics Portal*. Retrieved from: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[66] Statista. (2015b). Active reach of social networking services in the United Kingdom (UK) in March 2014, by device. *Statista: The Statistics Portal*. Retrieved from: http://www.statista.com/statistics/272805/social-networking-sites-active-reach-in-the-united-kingdom-by-device/

[67] Thomases, H. (2010). *Twitter Marketing: An Hour A Day*. Indianapolis: Wiley Publishing.

[68] Thompson, S. (1998). Why Ask Questions in Monologue? Language choice at work in scientific and linguistic talk. In: Hunston, S. ed. *British Studies in Applied Linguistics: Language at Work*. 137-150.

[69] Thornborrow, J & Wareing, S. (1998). *Patterns in Language: An introduction to language and literary style*. London: Routledge.

[70] Toth, E.L. & Heath, R. (2009). *Rhetorical and Critical Approaches to Public Relations*. Abingdon: Routledge.

[71] trendwatching. (2014). The Future of Customer Service. Retrieved from: `http://trendwatching.com/x/wp-content/uploads/2014/09/2014-09-FUTURE-OF-CUSTOMER-SERVICE.pdf`

[72] Twitter. (2014). The Twitter Glossary. Retrieved from: `https://support.twitter.com/articles/166337-the-twitter-glossary`

[73] Twitter Basics. (2015). Learn Twitter: How your business can use Twitter. Retrieved from: `https://business.twitter.com/basics/learn-twitter`

[74] United Biscuits. (2014). About McVitie's. Retrieved from: `http://www.mcvities.co.uk/about`

[75] Waugh, L.R. (2010). Pronominal choice in French conversational interaction: Indices of national identity in identity acts. In: Tanskanen, S-K., Helasvuo, M-L., Johansson, M. & Raitaniemi, M. *Discourses in Interaction*. Amsterdam: John Benjamins Publishing Co. 81-100.

[76] Yan, J. (2011). Social media in branding: Fulfilling a need. *Journal of Brand Management*. 18(9). 688-696. doi: 10.1057/bm.2011.19

[77] Zappavigna, M. (2012). *Discourse of Twitter and Social Media: How We Use Language To Create Affiliation On The Web*. London: Continuum. Retrieved from: `https://www.dawsonera.com`

The tweeting brand: When conversation leads to humanization

Andria Andriuzzi

Sorbonne Business School
Paris, France
andria@andriuzzi.com *

Abstract. Brands can now express themselves through conversation on social media sites such as Twitter but should marketers mix brand personification and conversational techniques together? Based on a series of interviews with web users, this research shows that brands conversing in this way can be seen to be more human and can improve their image even without the physical presence of a speaker. However, simply being present on Twitter or personifying the brand will not suffice. To be more human, brands must participate in quality conversation. Consumers evaluate conversation according to its context, content and form.

Keywords: Conversation, Brand, Personification, Anthropomorphism, Social Media, Twitter

1 Introduction

As "semiotic instances" (Semprini, 1992), brands have a long tradition of co-producing meanings. Brands convey their own values using advertising to shape people's perceptions across a number of years (Michel, 2013). With the development of the Internet, consumers are now considered to be fully involved in a co-creation process (Cova & Cova, 2009). Today, echoing the consumers power of expression, brands can also make their voices heard through conversation on social media sites like Twitter. As some consumers supposedly view brands as a 'person' thanks to anthropomorphism (Aaker & Fournier, 1995), the use of human language in interaction could generate such reactions. Alongside other personification techniques like the use of celebrities or fictional characters (Cohen, 2014), conversation can be seen as an attempt to humanize the brand by adopting a human behavior.

Just as certain everyday conversations can have a positive or a negative effect on us, we believe that every brand conversation has a different value. While several studies focus primarily on the number of comments or on their valence (de Vries, Gensler & Leeflang, 2012) and methodologies used by specialized agencies are sometimes not very clear (Tissier-Desbordes & Vernette, 2012), this research looks more specifically at the quality of such

* The author would like to thank Professor Géraldine Michel, Director of Chaire Marques & Valeurs, for her kind advice, Elise Yarwood, for her help in editing the article, and Xavier Cazard of Entrecom for his support. Correspondence concerning this article should be addressed to Andria Andriuzzi, IAE de Paris, 21 rue Broca, 75240 Paris Cedex 05, France.

conversation. The content analysis of a series of interviews reveals that a brand's behavior on social media sites can make it appear more human. However, what is more important than brand personification is the quality of conversation, creating an effect of anthropomorphism that positively affects the brand's image as well as word of mouth and purchase intentions. This article presents the three pillars of quality brand conversation from the consumers' point of view and it is, to our knowledge, the first instance in which these notions have been documented. Whether they are participating in brand conversations or simply observing, consumers evaluate the quality of brand conversation according to its context, content and form. Practitioners are therefore encouraged to develop personalized and valuable conversations in order to strengthen the links between brands and consumers.

2 Theoretical framework: Anthropomorphism and brand personification

Brands are not welcome on social media as these tools were originally created for human beings (Fournier & Avery, 2011). Nevertheless, they can be somewhat successful if they adopt the codes of social media sites, such as transparency and accepting a certain loss of control (ibid.). For example, community managers frequently converse in a way similar to that of other users on social media sites. Nevertheless, one study shows that for less-known brands, the adoption of an informal and conversational communication style counter-intuitively reduces confidence in the brand (Gretry et al., 2014). With this in mind, is it really a good idea for these brands try to look more human on social media sites like Twitter?

Anthropomorphism is the propensity of some people to attribute human characteristics to inanimate objects. It is an individual psychological characteristic (Epley, Waytz & Cacioppo, 2007). For many years, research has linked the concept of anthropomorphism to the relationship between brands and consumers (Aaker & Fournier, 1995). Anthropomorphism strengthens the relationship with a brand to the point that it can instill loving feelings (Rauschnabel & Ahuvia, 2014). While many articles deal with brand relationship (Fournier, 1998) and brand personality (Aaker, 1997), some researchers question their theoretical foundations and operational applications (Avis, 2012; 2013). Although the propensity of consumers to see brands as people is still debated, practitioners often try to make their brands look more human with the use of personification techniques. Brand personification is the use of a patronymic name, or the promotion of a celebrity, a fictional character or a brand employee in their external communications (Cohen, 2014). It can lead to users identifying with the brand as a fellow person, causing positive effects, particularly on attitudes towards the brand. It can, however, also lead to negative effects. For example, the use of employees in commercials is, strangely enough, something that can lead people to doubt the brand's sincerity (Fleck, Michel & Zeitoun, 2014). Similarly, suspicious consumers appear to have less trust in products with a patronymic brand name (Eskine & Locander, 2014). Moreover, a brand does not need to be personified for consumers to attribute human qualities to it. Logos can, in this way, portray a sense of personality that influences purchase intent (Payne et al., 2013) and the use of narrative themes inspired by history or folklore can humanize a brand (Hede & Watne, 2013).

A new aspect introduced by social media sites such as Twitter is the ability for brands to converse with consumers (Gensler & Völckner, 2013). When brands use conversation - a human characteristic if any - can it create the effects of anthropomorphism for the public online? When the "the brand is speaking" on Twitter, do consumers see the brand as a person? If they are perceived as more human because of these interactions, can it lead to positive attitudes from consumers? In order to answer these questions we carried out a qualitative study and interviewed consumers directly.

3 Method

We conducted 12 semi-structured, individual, face-to-face interviews of varied durations, lasting from fifty-eight minutes to one hour thirty-four minutes. This method is best suited to the study of motivations, attitudes and perceptions (Bardin, 1977). Additionally, this method was used to gain insights into the perceptions of Twitter users who actively participate in brand conversation but also those who simply view these interactions, as this category represents the majority of consumers. In general, most consumers interact very little with brands online (Mathwick, 2002). It is, however, interesting to investigate the perceptions of those that view these conversations but do not interact. With this in mind, these points of view were also taken into consideration. We used a convenience sampling strategy that was limited to people using the Internet and social media but did not filter these users according to their level of online interaction with brands. Furthermore, interviewees were not asked if they were Twitter users or not. This decision was made considering that some tweets or some Twitter conversations can be seen elsewhere other than on Twitter, if they are quoted in the media for example and we wanted to collected opinions and feelings from Twitter users as well as from non- users. Thus, we believe that to understand the impact of Twitter on consumers it is important to take into account the opinions of non-Twitter users as they can be exposed to the brand conversation vicariously. Considering that demographic characteristics can result in diverse reactions to marketing campaigns on social media (Campbell, Ferraro & Sands, 2014), we used age and gender as selection criteria. Two men and two women were selected from three age groups: 18-34, 35-59 and over 60. We also paid attention to the diversity of socio-professional categories and selected participants from a broad range of professions including: an engineer, an entrepreneur, a marine officer, a human resource manager, a painter, two public employees, a touristic guide, a librarian and 3 retired people - two former managers and a former assistant. An interview guide was used that included topics covering the presence of brands on social media, the observation of exchanges between brands and users and the possibility of participation in these conversations.

Even though it was not a selection criteria, we tried to understand if the interviewees had taken part in any Twitter brand interactions. That is to say we 'tried' to understand because it was not obvious that all interviewees had noted whether they had ever participated in "brand conversations" or even observed one. For example, it is interesting to note than during a whole one-hour interview, Denis, 68, denied ever having interacted with a brand online. Nonetheless, at the very end of the interview he remembered and thoroughly explained an interaction he had had with his bank, concerning a complaint that was posted on Twitter.

Among the twelve interviewees, nine of them observed some kind of brand conversation on social media, and three of them more specifically on Twitter. Half of them noted participating in brand interaction – even if this was as little as a simple comment on a brand post. This could be seen as a low number but it is, in fact, consistent with the rather low figures noted elsewhere concerning Twitter users participation, which is surprising considering the social media networks high level of awareness and media exposure. Indeed, research institute IPSOS stated that only 5 percent of the French population hold and use a Twitter account (2013). Furthermore, only 2 percent of web users report having ever commented on a brand post on social media, according to IFOP (2011). All of this considered, we were also interested in the reactions of people who observe without participating and we therefore presented interviewees with four stimuli in the form of screenshots at the end of each interview to observe how they reacted to the different types of online interactions between brands and consumers. Consequently, the twelve interviewees were able to express what they felt when observing Twitter brand conversations.

The transcripts of these interviews are 266 pages and 93,655 words long. After a dozen interviews, a phenomenon of information saturation was observed (Glaser & Strauss, 1967). Therefore, we deemed twelve interviews sufficient for the purposes of the study and went on to analyze our data. The collected material was subjected to a content analysis, as this technique can shed light on the non-explicit meaning of speech (Frisch, 1999). We made a preliminary analysis on a quick reading of the transcript and then used qualitative analysis software NVivo to code and categorize the meaning units in the order of utterance, organizing categories and subcategories throughout the analysis.

4 Results

With the analysis of complete and direct verbal contributions from interviewed participants (reproduced verbatim in this section), we are able to assess how web users perceptions vary depending on the nature of a speaker within brand conversation. We go on to discuss the positive and negative effects of conversation and to identify four key perceptions of brands depending on the representation of the speaker and the attitude of the users towards conversation. Subsequently, we present the criteria for quality conversation.

4.1 Conversational brands and anthropomorphism

When "the brand speaks". On Twitter, brands are often identified by their name and logo. However, it is possible that the source of the account is seen to be uncertain. How the source is perceived and represented mentally also varies significantly depending on the situation and on the Internet users. The "tweeting brand" generates different projections: the situation and the discourse imply the nature of the spokesperson, even if this person is not physically represented. The impression of a 'real conversation', if any, is created by the use of message personalization and the visibility of a brand's responses to other web users. Consumers can then assign human qualities to the brand such as a sense of empathy, someone that listens and responds:

> *"Yeah, I like it when it's an exchange, when it isn't just a claim from the client, but a claim which is echoed, and a response. I'm interested in interaction, in communication (. . .). They try to reach people just like a friend writing and responding. It's as if it's a friend. It's a communication style."* (Judith, 69).

Even without the physical representation of the speaker, a brand can appear to have a personality (Aaker, 1997). However, message personalization does not guarantee such a perception. For example, looking at an interaction between a brand and a user on Twitter, one interviewee doubted the sincerity of the exchange:

> *"I think that if a brand comes and talks to me specifically, then it's okay. But if a brand decides to talk to people it will be a sea, an ocean of comments and we'd never have time to read all the comments in the end. (. . .) I would think that it's an automated thing (laughs). Yes, like a machine. (. . .) A large chain wouldn't make the time for us . . . "* (Marina, 38).

Thus, the act of having a Twitter account is not sufficient to make a brand more human. The interactions on this account create different perceptions of the brand.

When the brand is personified. In some situations brands are represented on social media by a spokesperson, for example when experts answer on Twitter for consumer-care services. This is a personification technique comparable to the use of employees or celebrities in commercials (Fleck, Michel & Zeitoun, 2014). This human presence is often perceived as "more credible". However, although a number of the verbal responses collected indicate that this sort of brand representative can give the impression that the brand-consumer relationship is more human, it is not necessarily enough to gain the trust of consumers:

> *"Interactions with such delegates have to be . . . You must have the feeling. . . the feeling you really talk to someone, you know. Not with someone who doesn't care, not an automated response. You see. Because, well, you can also have. . . You can also talk to a real person who sounds like they are giving you an automated response. That's frustrating too. I like when you really have the feeling that you're speaking to a human being, that the guy's going to . . . even maybe add a touch of humor, something like that . . . You know, like in a real discussion. It's more reassuring."* (Guillaume, 39).

On social media, an interaction with a real person can be perceived as dehumanized. This perception reminds us of call centers, where the standardization of responses sometimes dehumanizes the relationship (Fielding, 2003).

To summarize, brands that speak on social media may seem more human, even without the physical representation of a speaker. Furthermore, personification does not guarantee an anthropomorphic effect on Internet users and brand behavior is, essentially, more important than the representation of a speaker. This leads us to consider the positive and negative effects of brand conversation.

4.2 Brand conversation effects on consumers

Positive effects of conversation. When brands are perceived as altruistic, authentic and caring, their attitude in conversation is considered as positive. Brand conversation has a number of potential effects on consumers and our interviewees noted that some conversations could influence purchase intents. This influence could come from the other web users participating in conversations, confirming the impact of online consumer-to-consumer communication (Carl, 2006). However, the participation of the brand can also play a role. In this way, brand conversation also influences attitudes towards the brand, especially for those who simply view the discussions, rather than participating. Furthermore, these conversations can generate word of mouth, both online and in everyday social contexts:

> *"In my opinion, a conversation online leads to another conversation outside. (. . .) If I hear of something like that, about a review, about a brand, a "maison", a product, well I want to have a look too. It makes you want to see if it's true, if it's wrong."* (Marc, 78).

From this we can see that exchanges between brands and consumers have effects comparable to online exchanges between consumers. Indeed, online word of mouth leads to people searching for more information and influences purchasing decisions (Keller, 2007). However, the behavior of brands can change the game, creating a positive impact as discussed above or resulting in negative attitudes towards the brand, which we will now consider.

Negative effects of conversation. Conversations are negatively evaluated when the exchanges are not personalized, unclear or are deemed unnecessary and unhelpful. In many cases, a brand's interventions are perceived as intrusive or manipulative. For example, brands can be suspected of using online interaction as a sneaky way to advertise themselves further.

> *"I'm not interested (. . .) If I buy a brand it isn't, well. . . I don't want. . . Well I do choose the brand; it isn't the brand that will overwhelm me with . . . well, you see. (. . .) There are too many reminders and all. Promotions, stuff like that. (. . .) You have those things, notifications you know, things like "do this, do that". And, so you see the. . . special offers, 50% off, blahblahblah. Sometimes you get two, three times the same thing, you know, almost one just after the other."* (Alexis, 42).

In many situations, there is simply no conversation: no one responds to brands or brands do not respond to comments. We found that conversational practices can also generate negative effects, specifically degrading the brand image. However, rather than its personification, it is the value of the conversation that is decisive.

4.3 Four perceptions of conversational brands

Table 1 shows four situations, highlighted by the analysis of the interviews, depending on the nature of the brand speaker and the attitude of web-users towards brand conversation. This is, to our knowledge, the first time that two of these situations have been described in

Table 1: Perception of brands depending on speaker and attitude toward conversation

	Brand Speaker	
Attitude towards conversation	**a brand**	**a spokesperson**
Positive	(1) Humanization	(3) Incarnation
Negative	(2) Standardization	(4) Bureaucratization

literature: a brand that is humanized due to the positive evaluation of its conversation and a spokesperson that is 'bureaucratized' because of the negative evaluation of the conversation she or he participates in.

(1) *The humanized brand*: When "the brand talks", it can make the brand seem 'more human' so long as the conversation appears to be human-driven. A non-personification of the speaker, represented in the form of a logo, still offers human qualities if its messages are personalized and oriented towards consumer satisfaction. It is, therefore, not necessary that an individual is represented for the brand to appear more human.
(2) *The standardized brand*: If the brand conversation is not personalized, one might think that a machine is tweeting or that the speaker does not wish to engage in a meaningful exchange. Just having a presence on Twitter, the simple use of the same tools as members of the public online is not enough to make a brand more personable.
(3) *The incarnated brand*: When a brand is represented by a specific speaker and that this speaker's conversation is deemed appropriate, this can lead to brand humanization as it creates the sense that there is a 'real discussion'. The human nature of the brand representative is clear and the users appreciate it. This strategy is similar to well-known personification techniques.
(4) *Bureaucratization*: Surprisingly, users may feel that a well-identified speaker is not animated by human feelings, or even is not human at all. When messages appear to be standardized and the conversation is not successful, it can lead to web users assuming that it is not, in fact, a brand representative responding, but an automated system. We found a paradoxical situation that implies that brand personification cannot always be achieved with the physical representation or name of a human being. To look more human, the brand must also portray human behavior. These findings can contribute to brand personification literature in that they go beyond the mere presence of a speaker and consider the value of its discourse. One interpretation of these results could be that success depends on the quality of brand conversation.

4.4 A quality conversation

The positive effects of a brand conversation can be explained by its perceived quality. An in-depth analysis of the interviews showed that the three pillars of quality brand conversation are (1) context, (2) content and (3) form. (1) It is most beneficial that conversations have a limited number of participants, so that they are easy to understand and follow. In a

quality conversation, participants are strongly involved and pay close attention to others. The conversation could have a utilitarian aspect, for example where a participants technical problem is solved either by the brand itself or by other Internet users. In this case, what the user needs is a rapid, relevant and useful response leading to a clear result and even if this is sometimes unexpected, it is always appreciated. However, whether to gain information or provide it, contributors seem to expect nothing in return. Thus, the value of the conversation lies in the quality of the interpersonal relationships that it creates. However, value can also be influenced by content. (2) Quality conversation must be, above all, interesting. Being able to learn something about the brand or from other users shows the intrinsic value of the conversation. Interviewees stated that they sometimes like to consult these interactions, commenting on their playful and hedonistic character. These results confirm that people react better to social media marketing operations when they are motivated to search for information and when they are entertained (Campbell, Ferraro & Sands, 2014). (3) The form of the conversation appears to play an equally important role in its value; the richness and precision of the exchange and the attention paid to the contents of the messages are particularly appreciated as is the inclusion of humor:

> *"I like to joke, playing with words and all... Yes, yes... In fact it is the... These conversations that attract me, that interest me, often things that are a bit funny, with a humorous side, or wordplays, you see. It makes me aware of... Otherwise if it's flat... No it's the humoristic side that attracts my attention... Yes, if it's fun, then you remember it better."* (Denis, 68).

To summarize, a quality conversation involves a small number of highly involved sincere and empathetic interlocutors. Apparently selfless, conversation is appreciated when it is interesting, useful or entertaining. Relevant and clear answers help to track the conversation until its conclusion. These results complement the existing methods used to measure the quality of the conversation by manually coding its relevance, timeliness, duration and frequency of messages (Adjei, Noble & Noble, 2010). Our analysis highlights the perceived quality of the conversation from a consumer point of view with three dimensions: context, content and form. It seems important for brand managers to take account of what makes a quality conversation on Twitter. We noticed that, when exposed to a conversation that they evaluate positively, consumers consider the brand to be more human, whether this is "a logo speaking" or whether a spokesperson is representing the brand. We also saw that brand conversation can have a positive effect on consumers, especially in terms of attitude towards the brand as well as the publicity gained by word of mouth. As one would expect, a consumer if much more likely to have a positive view of a conversation if they have enjoyed observing or participating in it. Nevertheless, our analyses help to better understand what makes a quality conversation. Furthermore, instead of linking positive effects to the valence or the number of comments, this research links them to the perceived quality of these interactions between brand and consumers.

5 Discussion and conclusion

In this research we have identified the concept of quality brand conversation, a notion that has not, to our knowledge, been previously identified. On the one hand, we have identified situations where the anthropomorphism of a brand has positive effects and a brand appears human through the use of quality conversation; on the other hand, we have discussed situations in which personification has negative effects, such as when human representatives partake in dehumanized conversation. Some of the situations described confirm theories suggesting that consumers attribute personality traits to brands (Aaker, 1997). However, we can neither confirm that a personality exists before an interaction with the brand, nor that this interaction guarantees that people think the brand has a personality. It is the quality of the conversation that helps to create anthropomorphism, even without the identification of a human speaker.

From a managerial point of view, these results could enable practitioners to rethink their brand presence on social media sites such as Twitter on three levels. (1) Users do not always want to have a conversation with brands. Therefore, brands should not necessarily be interested in developing interactions in which the sole purpose would be to increase the visibility of messages as this approach could harm their image. (2) However, brand conversation is appreciated if it brings value to users. The establishment of true customer relationship services on Twitter can, for example, helps to build consumer loyalty. (3) Brands can also benefit from developing conversational strategies to improve the quality of the conversation itself by running initiatives such as co-creation platforms. A quality conversation can generate useful ideas for the brand while improving its image. The measurement of the quality of the conversation - so far addressed through external indicators - could be the subject of further studies involving consumers directly.

To conclude, this research shows that although the human representation of the speaker may sometimes be seen as more credible, what consumers appreciate above all is brands that contribute to quality conversations, whether they are personified or not.

Bibliography

[1] Aaker, J., & Fournier, S. (1995). A brand as a character, a partner and a person: three perspectives on the question of brand personality, *Advances in Consumer Research*, 22, 1, 391-395.

[2] Aaker, J. (1997). Dimensions of brand personality, *Journal of Marketing Research*, 34, 347.

[3] Adjei, M., Noble, S., & Noble, C. (2010). The influence of C2C communications in online brand communities on customer purchase behavior, *Journal of the Academy of Marketing Science*, 38, 5, 634-653.

[4] Avis, M. (2012). Brand personality factor based models: A critical review, *Australasian Marketing Journal*, 20, 1, 89-96.

[5] Avis, M. (2013). *Humanlike brands and metaphor: applications and consequences,* Thesis, University of Otago.

[6] Bardin, L. (1977). *L'analyse de contenu*, Paris, Presses universitaires de France.

[7] Campbell, C., Ferraro, C., & Sands, S. (2014). Segmenting consumer reactions to social network marketing, *European Journal of Marketing*, 48, 3, 432-452.

[8] Carl, W. J. (2006). What's all the buzz about? Everyday communication and the relational basis of word-of-mouth and buzz marketing practices, *Management Communication Quarterly*, 19, 4, 601-634.

[9] Cohen, R. J. (2014). Brand personification: introduction and overview. *Psychology & Marketing*, 31, 1, 1-30.

[10] Cova, B., & Cova, V. (2009). Les figures du nouveau consommateur : une genèse de la gouvernementalité du consommateur, *Recherche et applications en marketing*, 24, 3, 81-100.

[11] De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. *Journal of Interactive Marketing*, 26, 2, 83-91

[12] Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, 114, 4, 864-886.

[13] Eskine, K. J., & Locander, W. H. (2014). A name you can trust? Personification effects are influenced by beliefs about company values, *Psychology & Marketing*, 31, 48-53.

[14] Fielding, G. (2003). Taking conversation seriously: The role of the call centre in the organisation's customer contact strategy, *Interactive Marketing*, 4, 3, 257-266.

[15] Fleck, N., Michel, G., & Zeitoun, V. (2014). Brand personification through the use of spokespeople: an exploratory study of ordinary employees, CEOs and celebrities featured in advertising, *Psychology & Marketing*, 31, 84-92.

[16] Fournier, S. (1998). Consumers and their brands: developing relationship theory in consumer research, *Journal of Consumer Research*, 24, 4, 343-353.

[17] Fournier, S., & Avery, J. (2011). The uninvited brand, *Business Horizons*, 54, 3, 193-207.

[18] Frisch, F. (1999). *Les études qualitatives*, Paris, Editions d'organisation.

[19] Gensler, S., & Völckner, F. (2013). Managing brands in the social media environment, *Journal of Interactive Marketing*, 27, 4, 242-256.

[20] Glaser, B.G., & Strauss, A.L. (1967). The Discovery of Grounded Theory. *Strategies for Qualitative Research*, Chicago, Aldine.

[21] Gretry, A., Horváth, C., van Riel, A., & Belei, N. (2014). "Don't you pretend to be my friend!" - The effect of brands' communication style on brand trust in brand-based online communities, working paper, Radboud University, Nijmegen School of Management.

[22] Hede, A.-M., & Watne, T. (2013). Leveraging the human side of the brand using a sense of place: case studies of craft breweries, *Journal of Marketing Management*, 29, 1-2, 207-224.

[23] Keller, E. (2007). Unleashing the power of word of mouth: creating brand advocacy to drive growth, *Journal of Advertising Research*, 47, 4, 448-452.

[24] Mathwick, C. (2002). Understanding the online consumer: A typology of online relational norms and behavior. *Journal of Interactive Marketing*, 16, 1, 40–55.

[25] Michel, G. (2013). *Management transversal de la marque: Une exploration au cœur des marques*, Paris, Dunod.

[26] Payne, C. R., Hyman, M. R., Niculescu, M., & Huhmann, B. (2013). Anthropomorphic responses to new-to-market logos, *Journal of Marketing Management*, 29, 1-2, 122-140.

[27] Rauschnabel, P. A., & Ahuvia, A. C. (2014). You're So Loveable: Anthropomorphism and Brand Love. *Journal of Brand Management*, 21(5), 1–39.

[28] Semprini, A. (1992). *Le marketing de la marque*, Paris, Editions Liaisons.

[29] Tissier-Desbordes, E., & Vernette, E. (2012). Le repérage marketing des influenceurs dans les réseaux sociaux : des dangers de l'ignorance aux risques de l'à peu près, *Décisions Marketing*, 67, 5-7.