GdPO Global Drug Policy Observatory
reportingmonitoringanalysis

Swansea University
Prifysgol Abertawe

GDPO Situation Analysis

June 2017

Corpus Linguistics Methodology on the Silk Road(s): The Escrow Example

Matteo Di Cristofaro[*]& Martin Horton-Eddison[¥]

Subject: The application of Corpus Linguistics to the concept of innovation in crypto-drug markets

This Situation Analysis is a methodological paper intended to supplement the research findings presented in GDPO Policy Brief 11, Horton-Eddison, M. & Di Cristofaro, M., *Hard Interventions and Innovation in Crypto-Drug Markets.*[1]

This analysis shows the application of Corpus Linguistics methodology and of CADS (Corpus Assisted Discourse Studies) approach to data extracted from Crypto-Drug Market (CDM) communities. The aim is twofold: First, to demonstrate how the analysis of textual data created by CDM users can help pin-point the impact that 'real-life' events (in this case, the FBI's seizure and closure of Silk Road, and the theft of a substantial amount of Bitcoin on Silk Road 2) have on online crypto-communities. Second, to illustrate how linguistics theories and methodologies may be used to investigate how 'trust' is established/reinforced in online CDM communities.

For the purpose of this study we have used static copies of two CDM forums: The original Silk Road, and Silk Road 2.0 (from now on, SR1 and SR2 respectively[2]). Each copy was collected by Gwern,[3] a few days before the seizures of the SR1 and SR2 servers on 3rd November 2013, and 4th November 2014, respectively. It must be noted that these snapshots – which represent the most complete copies to date – are incomplete 1:1 replicas of the websites as they were while online. Due to a series of technical restrictions - such as limited access permissions, and website downtimes - the snapshots contain only content that was accessible and online at the time of the crawling. The original data crawl by Gwern is thus acknowledged as incomplete:

[*] Independent Post-doctoral Researcher

[¥] PhD Researcher, Global Drug Policy Observatory

[1] Horton-Eddison, M. & Di Cristofaro, M, *Hard Interventions and Innovation in Crypto-Drug Markets: The escrow example*, GDPO Policy Brief No.11, GDPO, Swansea, June, 2017

[2] Silk Road 1 operated between February 2011 and 2nd October 2013, and Silk Road 2 operated between 6th November 2013 and 6th November 2014.

[3] Data mirrored by Gwern Branwen (pseudonym), available in raw form here: https://www.gwern.net/DNM%20archives

> […] any analysis must take seriously the incompleteness of each crawl and the fact that there is a lot and always will be a lot of missing data, and do things like focus on what can be inferred from "random" sampling […][4]

Accordingly, every analysis – including this one – must take into account the aforementioned limitations, and is consequently limited *a-priori*. The adoption of a methodology (Corpus Linguistics) and an approach (CADS) that allows the interrogation of the whole of the selected CDM archives – as one holistic database (i.e. a corpus) - is our approach to these limitations. By looking at the 'bigger picture' as a way of filtering and pin-pointing which parts of the data show peculiarities - and that may be further analysed quantitatively – we aimed to limit the presence of errors and/or inconsistencies.

For the purpose of our analysis we have focussed on those parts of the data (SR1 and SR2 snapshots) that include the forum messages, excluding those pages (e.g. user profiles and website statistics) that do not contain exchanges of messages between users. Among the data that was excluded are the so-called 'vendor pages': generally short profiles written by the vendors themselves to introduce their shop and their products, and to provide information regarding payments and shipping. These pages are also the ones where users can leave their feedback after a transaction has been made. Vendor pages were excluded for three main reasons: i) as we are conducting a linguistic analysis we need to collect as much data as possible. Vendor pages represent only a small part of the pages crawled[5] and contain small parts of (mostly descriptive) texts. ii) Linked to point i., feedback on vendor pages tends to be considered by the CDM users as "useless" because vendors are known to post fake reviews to increase their overall-rating[6]. iii) Even when present, textual data in vendor pages is not interactional.

Data extraction/formatting

Once the non-forum pages were filtered out, a combination of X-Path strings and of custom Python scripts were used to extract all the textual data contained in the forum pages, while preserving a layer of metadata for filtering/limiting the analysis on the basis of details such as: username of the author; publication date; title of the post; and name of the section in which the messages was posted in. Succinctly, the forum pages were first "converted" into pseudo-XML files, which were then used to feed a relational database. This resulted in an annotated corpus for each website; each corpus was then loaded into CQPweb, a corpus analysis framework whose purpose is to conduct linguistic analysis.[7]

Methodology

The employment of CADS in digital media research is well attested, and has supported analyses of e.g. influence, ideology, immigration, and social benefits.[8] CADS relies on the integration of a quantitative-oriented methodology (Corpus Linguistics) and a qualitative-focussed approach (Discourse Studies), which facilitates the task of interpreting discourses and attitudes in vast datasets. The researcher/s can therefore rely on statistical methods to 'pinpoint[...] areas of interest for a subsequent close analysis.'[9] The approach is typically inductive, 'hovering between the corroboration of what is felt to be known' and "serendipitous" discovery.[10] CADS requires

[4] http://www.gwern.net/DNM%20archives#interpreting-analyzing Accessed, 10th March 2017

[5] 1102 pages for SR1, 671 for SR2. For comparison, the forum pages included in the two corpora amount to 193,622 and 53,476 files for SR1 and SR2 respectively.

[6] Lorenzo-Dus and Di Cristofaro (in preparation) 'I know this whole market is based on the trust you put on me and I don't take that lightly': A Corpus Assisted Discourse Studies approach to trust in Silk Road.

[7] http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf

[8] See Baker et al 2013; Zappavigna 2013; Baker and McEnery 2015; Lorenzo-Dus and Di Cristofaro 2016; Prentice et al 2013

[9] See Baker et al 2008, P.28, Baker and McEnery (ed.) 2015

[10] See Partington 2006, P.12 & Marchi et al, 2017

a degree of engagement with the data[11] that is often grounded on extra-linguistic knowledge concerning the data itself and its authors (the virtual Silk Road communities in our case). This renders the methodology particularly fitting for inter-disciplinary research, where different background and knowledge levels work towards the same aim. The analysis of the SR1 and SR2 forums was therefore characterised by a constant interaction of theoretical frameworks (linguistics and public policy), language data (the SR1 and SR2 corpora), and extra-linguistics knowledge (technical and historical aspects of the two communities).

The quantitative analysis was largely conducted through the lens of *collocations*, a notion that relies on the knowledge that the relations existing between (and among) words create meanings. In other words (emphasis in the original)

> "[…] the term *collocation* denotes the idea that important aspects of the meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which it frequently co-occurs […]."[12]

Therefore, the meaning(s) that a combination of two words W1 and W2 conveys is merely not the product of "meaning W1" + "meaning W2", but it can be a meaning that is unique to this combination and only loosely based on the meanings of the single elements. This theoretical aspect is operationalised through the calculation of *collocates*, i.e. the words that significantly co-occur in the corpus with the searched word. A *collocate* is therefore the result of a quantitative analysis that, through statistical measurements, seeks to identify the relation between a word and the words that appear with it in the data, and to help the researcher identify features of the data that are both salient and peculiar.

In tables 1 and 2 we have included the top 25 collocates of our search term *escrow* in SR1 and SR2 respectively to illustrate our approach:

SEQ Illustration * ARABIC1. Illustration: SR1 escrow collocates

Rank	Collocate (lemma)	Observed collocate freq.	Log-Likelihood
1	in	11998	22717.892
2	stay	3076	18601.919
3	system	3051	16579.63
4	out	3168	5468.932
5	outside	843	4210.66
6	Sheep	915	3726.683
7	Hedge	508	3634.341
8	release	681	3424.318
9	full	898	3416.609
10	service	978	2996.513
11	no	2077	2669.195
12	escrow	675	2379.271
13	fund	547	2107.805
14	within	646	2067.155
15	of	5562	1933.389

[11] cf. Partington, Duguid, Taylor, 2013, Pp. 11-14

[12] McEnery and Hardie, 2012, Pp.122-123

16	deal	701	1684.941
17	money	892	1675.15
18	use	1764	1671.26
19	se	284	1630.471
20	through	835	1531.887
21	and	6869	1454.823
22	hold	456	1354.734
23	transaction	585	1300.596
24	protection	258	1281.842
25	hedging	171	1278.128

2. Illustration: SR2 escrow collocates

Rank	Collocate (lemma)	Observed collocate freq.	Log-Likelihood
1	system	2918	17565.259
2	in	8524	13278.991
3	pend	1696	12486.737
4	no	2885	5359.098
5	full	1051	4245.852
6	fund	877	3857.497
7	money	1368	3252.235
8	Offer	894	3028.596
9	release	628	2944.594
10	stay	808	2767.16
11	balance	623	2673.937
12	centralized	281	2397.929
13	without	767	1985.973
14	implement	401	1857.834
15	Pending	232	1811.04
16	with	2523	1325.092
17	tie	259	1322.685
18	use	1366	1251.426
19	an	1345	1161.093
20	stick	403	1148.826
21	coin	680	1131.867
22	market	678	1052.886
23	until	530	1047.498
24	there	1345	916.29
25	dispute	220	816.316

It must be noted here that the Log-Likelihood values cannot be compared across the two different corpora (i.e. SR1 and SR2), since Log-Likelihood is a statistical significance measure and its results are specifically dependent on the size of the data being analysed. Therefore e.g. the word *no* – which appears as collocate of *escrow* in both SR1 and SR2 – is ranked 11th in SR1 and 4th in SR2, with Log-Likelihood values of 2669.195 and 5359.098 respectively. These two Log-Likelihood values cannot be compared: it cannot be stated that e.g. *no* is a collocate whose significance to *escrow* is double in SR2 than it is in SR1. In fact, Log-Likelihood shows that the appearance of a collocate is not due to chance: that its relation to the search term is statistically significant.

It goes without saying that this quantitative approach is far from being sufficient in describing *how* escrow is used in the data, and to analyse the relation it has with its collocates. Hence these results have to be interpreted qualitatively, specifically through manual analysis of the term's occurrences (or of a sample of occurrences), where the collocates for the searched term (escrow) act as "entry points" to discover "how" escrow is discussed – and how the concept of escrow is *used* - in the two online communities under consideration.

Analysis

The aim of the qualitative analysis of the collocates calculated by means of the quantitative approach is to understand if – in the case of our case study – there has been a change in the way in which escrow is "written about." The importance of adopting a qualitative approach can be exemplified by describing what the analysis of those occurrences – in SR1 and SR2 – where *system* is a collocate of *escrow* show. The word *system* is a highly significant collocate in both corpora, but it is used in different ways – and with different connotations – in the two data sets. In SR1 it is used to refer to "positive" aspects of the *escrow*, which is understood as a way of protecting the users and as a barometer to evaluate other markets, as illustrated in examples (1) – (4):

(1) The escrow system is in place for a reason - to provide protection (SR1)

(2) Since i had registered on the BMR for some counterfeit money but i got scammed pretty bad so now I want to stay within so old trusted SR escrow system . (SR1)

(3) I swore I'd never go anywhere else because SR was the most sophisticated of all the drug sites and had wide array of quality products. And the fucking escrow system was fucking awesome. (SR1)

(4) They do have an escrow system and it's almost sad how many SR users are over in their forums writing post after post about how they won't buy from Sheep until an escrow system is implemented-- all because they heard from someone else that there is no escrow without investigating it for themselves. (SR1)

Conversely in SR2 *system* is used as a collocate of *escrow* in ways that denote a "negative" stance, based on the fact that the *escrow system* used in SR1 ultimately resulted in the loss of Bitcoin during the FBI's seizure of the site, and users were urging the market administrators to create a new *escrow system* in SR2 that would guarantee financial security, as examples (5) – (8) illustrate.

(5) Don't forget to use PGP and I would recommend not purchasing from the Silk Road [2.0] just yet, Wait for the administrators to implement their escrow system so as to avoid being scammed as much as possible. (SR2)

(6) I would not be surprised if it was somewhere in between Evolution's Multi-Sig Escrow System and Alpaca's Multi-Sig LITE Escrow System . It might be the right balance between functionality and ease of use … … I guess we will have to wait 7 days to see … … Unless of course Defcon decided to release the Escrow System a few days early. (SR2)

(7) Defcon Defcon, when are you going to make an announcement on The Multi-Sig Escrow System? (SR2)

(8) The last attack that happened - in which all the coins were stolen - ultimately was because of the fact that silk-road had a centralized escrow system (in which the attacker allegedly exploited the 'check deposits' button and was apparently able to empty SR 's entire centralized escrow funds) . (SR2)

As exemplified in (8), one of the major concerns of SR2 users – in relation to *escrow* – are the consequences of remaining with a centralized system. This concern was also signalled by the presence of the collocate *centralized* (ranked 12th) in SR2; interestingly, neither *centralized* nor *decentralized* appeared as collocates of *escrow* in SR1. In SR2 *centralized* was used to denote a 'highly negative' connotation, particularly after the seizure of SR1 and the scam on SR2:

(9) Silk Road will never again be a centralized escrow storage. This week has shown the collateral damage we can cause (SR2)

(10) From this point forward DO NOT trust markets with centralized escrow. Use multi-signature transactions whenever possible, with trusted third parties as escrow providers. (SR2)

(11) Statistically speaking, almost all centralized escrow markets eventually end up ripping off everyone for all the coins. (SR2)

(12) The marketplace will relaunch as no-escrow . We will not re-implement escrow unless it is multi-signature and decentralized to multiple escrow providers (trusted mediators with feedback just like vendors). Never buy from a market which uses centralized escrow again. You will only get hurt no matter how honest the team is. (SR2)

(13) Buyers: do not purchase using centralized escrow. Use markets which have implemented multi-signature, or only purchase with No-Escrow (FE) from VERY trusted vendors. (SR2)

(14) How do you know that they won't suddenly decide to take all of the coins like ALMOST EVERY OTHER MARKET with centralized escrow has done? I don't, but I'm not gonna just abandon everything out of fear ... (SR2)

Therefore, both positive and negative views towards *escrow* were identified in SR2 but - as noted by Horton-Eddison and Di Cristofaro[13]- in different time-periods. The analysis showed that an upwards shift occurred around 13th February 2014, coinciding with a major theft of Bitcoins (SR2b), but had nevertheless started earlier after the seizure of the first Silk Road (SR2a). The identification of these shifts was made possible by both the qualitative analysis and the adoption of filtering techniques (based on the messages' metadata) for the quantitative results. The data showed that the stance that users had towards *centralized escrow* in SR2 clashed with the one that users had in SR1; examples of the latter are provided by the collocate *protection* (SR1, rank 24th). The collocation *escrow* + *protection* had a 'positive' connotation in SR1, where users described the safety benefits that the *escrow* system provided:

(15) The EU vendors just seem so dodgy with weird things going on ie amsterdamshop, planta etc. North America has far superior buds at far superior prices but you don't get the protection of escrow or refunds that north american buyers get , that is understandable. (SR1)

(16) SR policy says that you should never FE, as the escrow system is protection for both you and the seller. (SR1)

(17) SR is a business, and they get no commission on bank transfers, not to mention you have no escrow protection with a bank transfer. (SR1)

(18) You do want the protection of escrow don't you? Haven't you seen how many scammers there are on here? (SR1)

The use of *protection* as collocate of *escrow* in the broader SR2 data (rank 31st) revealed an increasingly suspicious attitude on behalf of the users, who either condemned the 'old centralized system' or doubted the way in which SR2 administrators were dealing with 'escrow protection' in light of escrow's perceived failures:

[13] Horton-Eddison, M. & Di Cristofaro, M, *Hard Interventions and Innovation in Crypto-Drug Markets: The escrow example*, GDPO Policy Brief No.11, GDPO, Swansea, June, 2017

(19) This shows the fact that their is no protection via escrow for buyers right now. All escrow will do right now is hold the coins in limbo for 17 or whatever days it is. At that point those coins become the vendor's . IF SR [2.0] updates it's shit and actually starts assisting disputes, this changes things a bit - but currently that's not the case. (SR2)

(20) if your dealing with a trusted vendor FE is only going to make it impossible for the customer to try to scam FE'ing was never promoted on any marketplace and basically "voided the warranty" - your escrow protection was obviously lost and the administrators of the market have no responsibility in resolving any disputes. (SR2)

Summary

Our analysis has shown a way to approach large sets of textual data to understand how a given topic (*escrow* in this case) is discussed in online communities. This topic was used as a way to understand how financial transaction *trust* was built in the Silk Road and Silk Road 2.0 online communities. The data was collected by extracting the messages that were posted on the two websites forums, and was analysed by adopting the CADS methodology. This involved the adoption of quantitative (statistical) procedures as a way to funnel down salient aspects of the dataset, allowing the researchers to qualitatively interpret and understand vast amounts of language data (while avoiding the dangers of attributing qualitative values to quantitative results without further interpretation). As showcased in this paper (see Horton-Eddison and Di Cristofaro for details[14]), the methodology made it possible to compare the users' attitudes toward the use of an *escrow* system, and to identify a shift (from positive to negative attitude) as a consequence of an extra-linguistic events, such as law enforcement operations and the later Bitcoin theft. This paper has demonstrated how the CADS methodology can be applied to understand and interpret how their members "talk about" a specific topic by directly engaging the linguistics aspects and findings within a public policy framework. The interdisciplinarity of the methodology has led to the identification of precise real-life events which brought about sudden and inverse shifts in how the users of the community related to the topic under examination, and in so doing, helped identify innovation trends in the communities of study.

ACKNOWLEDGEMENTS

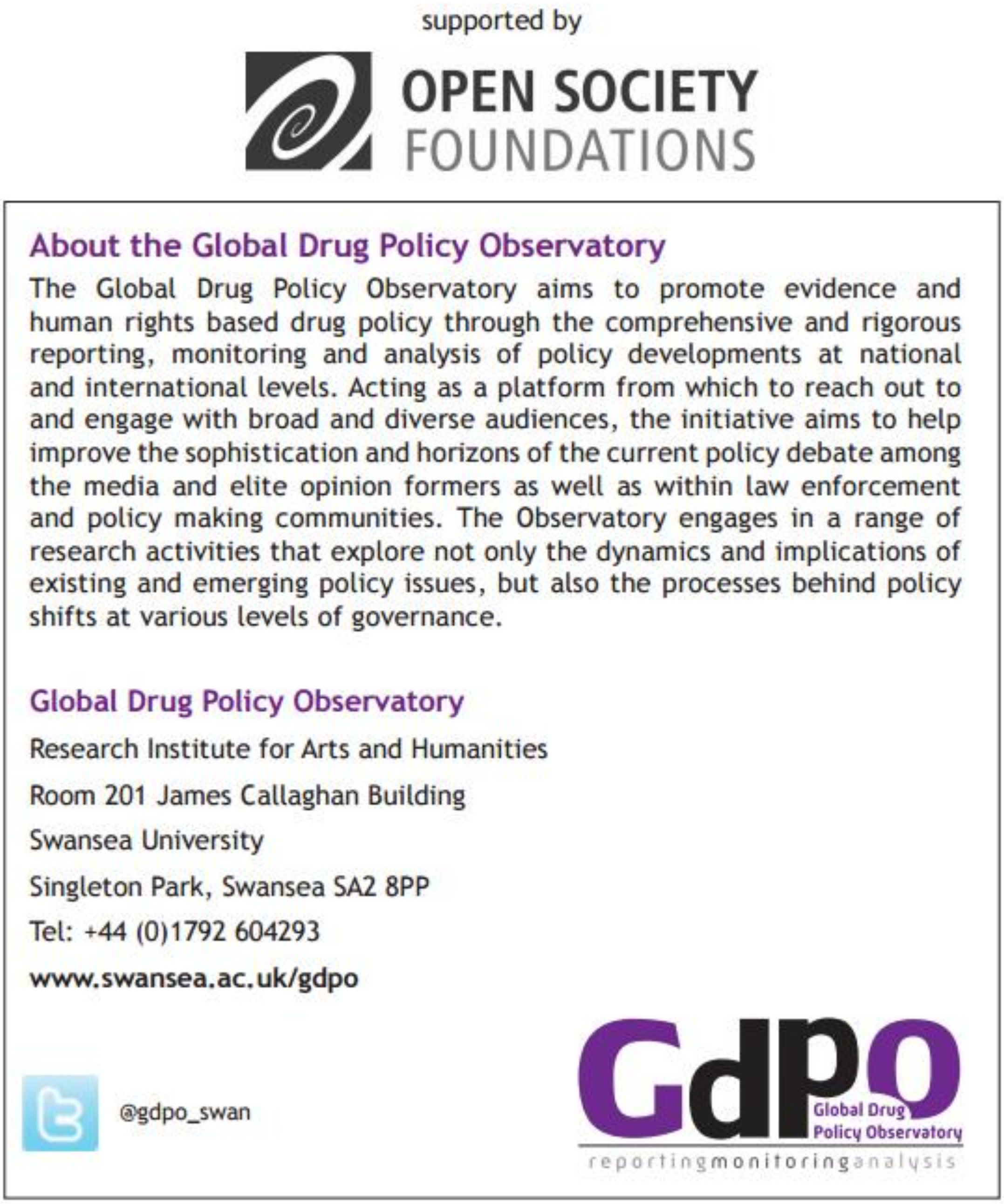
The authors would like to thank Professor Dave Bewley-Taylor, Director of the Global Drug Policy Observatory, Research Institute of Arts and Humanities, Swansea University, and Professor Nuria Lorenzo-Dus, English Language and Applied Linguistics, Swansea University for comments on earlier drafts of this Situation Analysis. Thanks also for insightful feedback and comments received from the participants of the roundtable on 'Cyber-Trust in Crypto-Drug Markets: Implications for Policy and Policing' held at the International Institute for Strategic Studies, 4th April, 2017.[15] The authors would also like to thank the CHERISH-DE initiative for supporting this research. Any errors of fact or interpretation remain with the authors.

cherishde

Swansea University
Prifysgol Abertawe

[14] Horton-Eddison, M. & Di Cristofaro, M, *Hard Interventions and Innovation in Crypto-Drug Markets: The escrow example*, GDPO Policy Brief No.11, GDPO, Swansea, June, 2017

[15] For an account of the event, see Juan Fernandez Ochoa 'Trust in the Crypto-Drug Markets' http://gdpo.swan.ac.uk/?p=466



supported by

OPEN SOCIETY FOUNDATIONS

About the Global Drug Policy Observatory

The Global Drug Policy Observatory aims to promote evidence and human rights based drug policy through the comprehensive and rigorous reporting, monitoring and analysis of policy developments at national and international levels. Acting as a platform from which to reach out to and engage with broad and diverse audiences, the initiative aims to help improve the sophistication and horizons of the current policy debate among the media and elite opinion formers as well as within law enforcement and policy making communities. The Observatory engages in a range of research activities that explore not only the dynamics and implications of existing and emerging policy issues, but also the processes behind policy shifts at various levels of governance.

Global Drug Policy Observatory

Research Institute for Arts and Humanities
Room 201 James Callaghan Building
Swansea University
Singleton Park, Swansea SA2 8PP
Tel: +44 (0)1792 604293
www.swansea.ac.uk/gdpo

@gdpo_swan

GdPO Global Drug Policy Observatory
reportingmonitoringanalysis