



**LOGIC OF SHARED SIGNIFICATIONS ON  
*INTERNET RELAY CHAT***

DAVID-OLIVIER MERCIER

Thesis submitted to the University of Ottawa  
in partial fulfillment of the requirements for the  
Master of Arts in Communication degree

Department of Communication  
Faculty of Arts  
University of Ottawa

© David-Olivier Mercier, Ottawa, Canada, 2019

**Abstract:** Through the observation of conversations on *Internet Relay Chat* and the quantitative analysis of “*chat-logs*”, I investigate the characteristics of this form of communication unique to the digital realm. My research rests on a theoretical framework integrating the *semiotic* and *pragmatism* of Charles S. Peirce (as primary groundwork) with the philosophy of Ludwig Wittgenstein and the sociology of Erving Goffman, to grasp shared significations in cyberspace simultaneously as logical process and as social practice. This exploratory case study yields evidence supporting the potential fruitfulness of Peircean philosophy as the foundation for a new paradigm in empirical communication research, and successfully puts to the test a particular type of method (computational and diagrammatic) suggested to accomplish such research.

**Résumé :** À travers l'observation de conversations sur l'*Internet Relay Chat* et l'analyse quantitative de “*chat-logs*”, j'investigue les caractéristiques de cette forme de communication unique au domaine numérique. Ma recherche repose sur un cadre théorique intégrant la *sémiotique* et le *pragmatisme* de Charles S. Peirce (comme fondement principal) avec la philosophie de Ludwig Wittgenstein et la sociologie d'Erving Goffman, pour saisir les significations partagées dans le cyberspace simultanément comme processus logique et comme pratique sociale. Cette étude de cas préliminaire apporte des éléments de preuve corroborant la fécondité potentielle de la philosophie Peircéenne comme fondement d'un nouveau paradigme de recherche empirique en communication, et met à l'épreuve avec succès un type particulier de méthode (computationnelle et diagrammatique) proposé pour accomplir de telles recherches.

**Keywords:** *semiotic, pragmatism, observation, quantitative analysis, IRC*

**Mots-clefs :** *sémiotique, pragmatisme, observation, analyse quantitative, IRC*



|                                                                   |        |
|-------------------------------------------------------------------|--------|
| 4. 3 (d) – Interpretation of Figure 6a and Figure 6b .....        | p. 97  |
| 4. 3 (e) – Combined analysis of all Ubuntu-related diagrams ..... | p. 101 |

## 5 – Conclusions

|                                                                                                                             |        |
|-----------------------------------------------------------------------------------------------------------------------------|--------|
| 5. 1 – Findings of Prior Research that Are Corroborated by this Work                                                        |        |
| 5. 1 (a) – The structure of IRC conversations mirrors the platform's technical attributes .....                             | p. 105 |
| 5. 1 (b) – The statistical analysis of IRC logs can reveal otherwise elusive characteristics .....                          | p. 108 |
| 5. 2 – Four Hypotheses Herein Put to the Test and Substantiated by Newfound Evidence                                        |        |
| 5. 2 (a) – Each form of semiosis has by-products which are isomorphic to a corresponding type of semiotic network .....     | p. 110 |
| 5. 2 (b) – Diagrams can reveal more aspects of complex semiotic patterns than other means to that end .....                 | p. 113 |
| 5. 2 (c) – This study's diagrams mirror how different “discourse communities”, or “language-games”, interpenetrate .....    | p. 115 |
| 5. 2 (d) – Established semiotic patterns can withstand sweeping changes, including ones in setting or in participants ..... | p. 118 |
| 5. 3 – Major Limitations of this Research                                                                                   |        |
| 5. 3 (a) – Lack of statistical representativeness .....                                                                     | p. 121 |
| 5. 3 (b) – Choice of settings for KH Coder’s diagrams .....                                                                 | p. 122 |
| 5. 3 (c) – Choice of stopwords .....                                                                                        | p. 124 |
| 5. 4 – Conceivable Practical Uses for the Findings Now Under Consideration .....                                            | p. 125 |
| 5. 5 – Overview of Paths for Further Research .....                                                                         | p. 127 |

## 6 – Primary Diagrammatic Representations

|                                                                           |        |
|---------------------------------------------------------------------------|--------|
| 6. 1 – Diagrammatic Representations of #libertyinmind and #politics ..... | p. 130 |
| 6. 2 – Diagrammatic Representations of Ubuntu-related channels .....      | p. 134 |

|                             |        |
|-----------------------------|--------|
| <u>7 – References</u> ..... | p. 142 |
|-----------------------------|--------|

## 8 – Appendixes

|                                                                                                |        |
|------------------------------------------------------------------------------------------------|--------|
| 8. 1 – <i>Appendix 1: Acknowledgements</i> .....                                               | p. 149 |
| 8. 2 – <i>Appendix 2: List of Publicly Logged IRC Channels Analyzed</i> .....                  | p. 150 |
| 8. 3 – <i>Appendix 3: Data Pre-Processing</i> .....                                            | p. 153 |
| 8. 4 – <i>Appendix 4: Various Perl Scripts Used for Pre-Processing</i> .....                   | p. 155 |
| 8. 5 – <i>Appendix 5: List of Stopwords Used</i> .....                                         | p. 157 |
| 8. 6 – <i>Appendix 6: KH Coder Interface Example</i> .....                                     | p. 160 |
| 8. 7 – <i>Appendix 7: Ludwig Wittgenstein: Philosophical Investigations</i> .....              | p. 163 |
| 8. 8 – <i>Appendix 8: Erving Goffman: The Presentation of Self in Everyday Life</i> .....      | p. 171 |
| 8. 9 – <i>Appendix 9: Additional Methodological Concerns</i>                                   |        |
| 8. 9 (a) – Content analysis & computational linguistics .....                                  | p. 179 |
| 8. 9 (b) – Methodological pansemioticism .....                                                 | p. 182 |
| 8. 9 (c) – Similarity or dissimilarity of channels studied .....                               | p. 183 |
| 8. 10 – <i>Appendix 10: Additional Diagrams – Wikimedia-related channels</i> .....             | p. 186 |
| 8. 10 (a) – Interpretation of Figure 7a and Figure 7b .....                                    | p. 188 |
| 8. 10 (b) – Interpretation of Figure 7c .....                                                  | p. 195 |
| 8. 10 (c) – Preliminary interpretation of Figure 8a and Figure 8b .....                        | p. 200 |
| 8. 10 (d) – Interpretation of Figure 8c .....                                                  | p. 203 |
| 8. 10 (e) – Involvement of “bots” .....                                                        | p. 204 |
| 8. 10 (f) – Log excerpts corresponding to Figure 9c .....                                      | p. 208 |
| 8. 11 – <i>Appendix 11: Additional Diagrams – Including Additional Verbs</i> .....             | p. 209 |
| 8. 12 – <i>Appendix 12: General Features of the Data Analyzed</i> .....                        | p. 219 |
| 8. 13 – <i>Appendix 13: Sample Table of Word Frequencies</i> .....                             | p. 223 |
| 8. 14 – <i>Appendix 14: Sample of Similarity Matrix (Jaccard Similarity Coefficient)</i> ..... | p. 224 |
| 8. 15 – <i>Appendix 15: IRC Interface Example (Quassel Client)</i> .....                       | p. 225 |
| 8. 16 – <i>Appendix 16: IRC Log Sample – Spam in #ubuntu-offtopic</i> .....                    | p. 226 |
| 8. 17 – <i>Appendix 17: Creative Use of Emoticons or “Smileys”</i> .....                       | p. 228 |

**Note on citations:**

The formatting employed for in-text references varies depending on whether or not the citation refers to works by Charles S. Peirce and Ludwig Wittgenstein. However, in all cases, the usual parentheses marking out references have been replaced by brackets. That is intended to better emphasize the references, to avoid confusion where many nested and concatenated parentheses would otherwise have occurred, and to facilitate the computational analysis of the text.

For references to each of the works by Peirce quoted in the present document, the notation most commonly used in Peirce scholarship was followed, on a case-by-case basis, as detailed notably by Vincent Michael Colapietro in his 1989 treatise *Peirce's Approach to the Self*. One list of such abbreviations, omitting that which is customary for unpublished manuscripts but otherwise ostensibly exhaustive, can be found on the following Wikipedia page:

[http://en.wikipedia.org/wiki/Charles\\_Sanders\\_Peirce\\_bibliography](http://en.wikipedia.org/wiki/Charles_Sanders_Peirce_bibliography)

Accordingly, for example:

- [CP 1.23] = *The Collected Papers of Charles Sanders Peirce*, volume 1, paragraph 23
- [SS 80—81] = *Semiotics and Significs: Letters to Lady Welby*, page 80–81
- [PPM 167] = *Pragmatism as a Principle and Method of Right Thinking*, page 167
- [MS [R] 491: 2] = Unpublished manuscript #491, page 2

A similar pattern, commonly found in Wittgenstein scholarship, was used for references to the works of that author. Thus, for example:

- [TLP 3.203] = *Tractatus Logico-Philosophicus*, proposition 3.203
- [PI §110] = *Philosophical Investigations*, paragraph 110

For all other references, the APA citation style was followed. When one author and two dates are present, the first date is the publication date of the first edition of the work while the second is that of the particular edition being quoted.

“*Thus my language is the sum total of myself; for the man is the thought.*” – Charles S. Peirce [CP 5.314]

“The limits of my language *mean the limits of my world.*” – Ludwig Wittgenstein [TLP: 5.6]

## **1 – Introduction**

### **1.1 – Personal Experience and Interest**

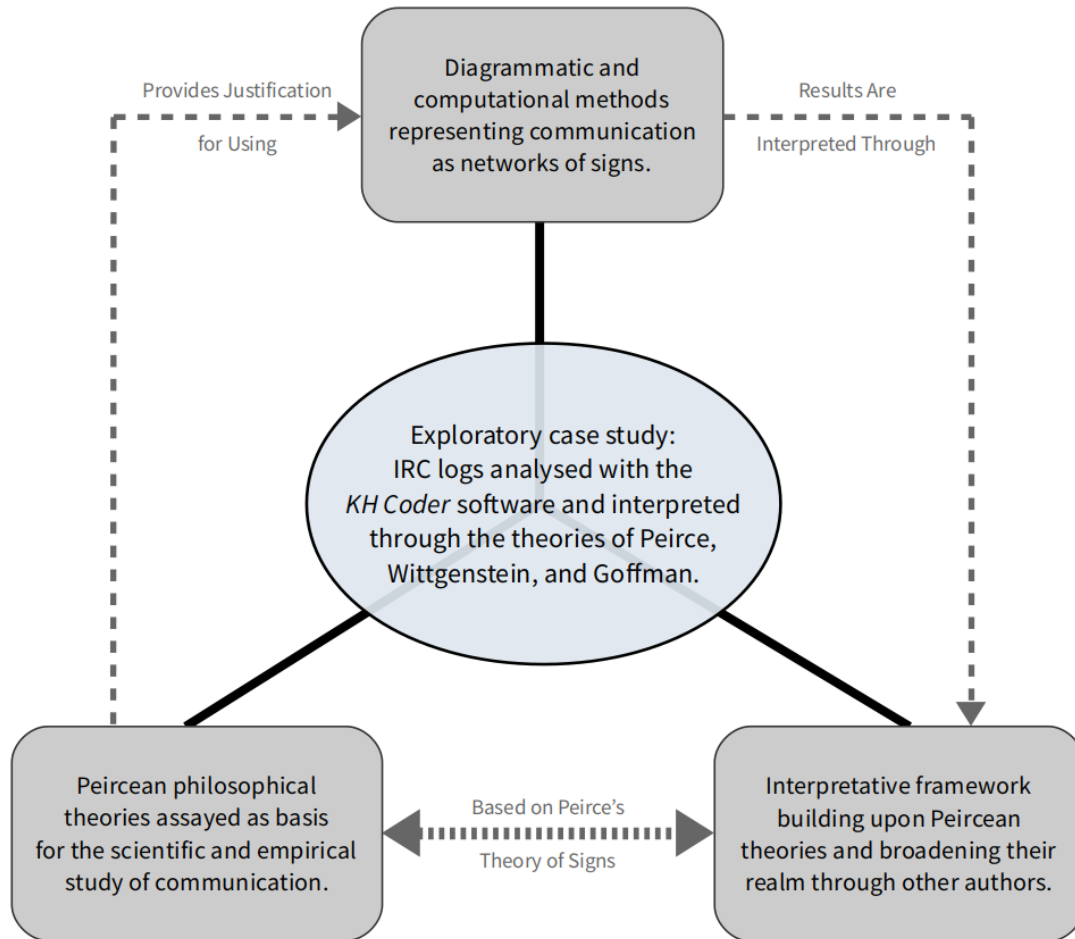
One personal interest of mine that can account for the subject matter of the most part of this thesis is my passion for cyberspace. Self-identifying as a *geek* from an early age and taking great pleasure in written communication, I have made nearly as many friends online as off-line. My experience with text-based communication through the Internet has given me keen awareness of the power of written language over even the very most important matters, and I have spent over half a decade setting up a non-profit organization (now called *Liberty in Mind*), initially exclusively and still<sup>1</sup> primarily through the written word, with a small crew of activists from around the world whom I had at first never met in person. As such events came to unfold in parallel with my studies, I became increasingly fascinated by the effects of digital media on individuals and on society at large.

Concurrently, I was disappointed to find that while my communication classes focused on how people interact with media, next to nothing was said of the nature of the building blocks of such interactions. In short, it was generally as if the process of communication had, when all is said and done, little to do with signifying or signs<sup>2</sup>. To address this shortcoming, I decided to study the digital media so dear to me from the perspective of *semiotic*, namely the study of signs and signification.

1 As of September 2019, I am on hiatus from the organization to focus on completing the present thesis.

2 Contemporary semiotician Umberto Eco declares that it is *common sense* to say that “any signification system is elaborated to produce communication processes” [1988, p. 29] (translation my own).

## 1. 2 – General Object of Inquiry and Aim of this Research



The diagram above represents the present work's *structure* and *purpose*. While some of its details will only become fully intelligible upon reading subsequent sections, its primary and most straightforward elements correspond to the fundamental building blocks of this research endeavour.

Accordingly, as indicated at the *bottom left* of the diagram, the starting point of this undertaking was the following inquiry: can the philosophy of Charles S. Peirce (1839—1914)<sup>3</sup> – primarily his two most lasting theoretical contributions: *semiotic* and *pragmatism* – serve as a fertile

3 I was introduced to the author through the insightful recommendation of Prof. Pierre Lévy, and later learned his theories under the skillful guidance of Prof. Fernando Andacht.

basis for the scientific and empirical study of actual instances of communication? If so, how? Even a cursory look at the literature reveals that the American philosopher's ideas, albeit popular, have seldom been employed in any such way, and with reasonable certainty *never* in a way genuinely approaching how I will herein suggest they be put to use<sup>4</sup>.

Thus, from a more practical point of view, the present research seeks primarily to substantiate the adequacy of Peircean philosophy as basis for conducting a scientific study of real-world (here, written and group) communication. For that purpose, it was necessary to determine, on the one hand, which methods would be most compatible with Peirce's thought, and on the other hand, how a complete analytical framework could be formulated by comparing and integrating Peircean concepts with some similar notions from other authors. This would, I inferred, ultimately enable me to devise an exploratory case study meant to provide a preliminary answer to my initial research question.

As will be later explained in detail, the synthesis of several pivotal Peircean concepts results in a conception of cognition and communication alike as inferential, virtual networks of signs, acting through the twin mechanics of representation and interpretation. The mind itself, according to Peirce, "is a sign resulting from inference" [CP 5.313]. Indeed, one's own train of thought can only realize such traits as "identity" and "personality" after one has learned to use peculiar sign-systems that "[owe] their origin [...] to human convention" [CP 5.532] – and thus to collective interpretation.

Most importantly, signs exhibit *law-like behaviour*: in characteristic manner, each sign brings about (probabilistically) regular consequences in the form of actual – or potential [CP 2.92] – signs entailed through interpretation. Acceptance of this premise is of crucial importance for the statistical

4 On that topic, see also [1.4 – Overview of the Theoretical Framework](#).

study of sign use: only if we assume that signs behave logically and regularly can we expect to find meaningful statistical patterns in actual instances of semiosis (sign-generation). Another important feature of semiosis is that it is determined by *final causation*, of which a particular category of special interest here is that of the *purposes* which guide human consciousness – itself “a special, and not a universal, accompaniment of mind” [CP 7.366]<sup>5</sup>.

A final aspect of Peirce's philosophy which I deemed at the outset to be of special importance for my aims is the unique place given to *mathematical reasoning*, and more generally to *diagrams* – as in some sense, all “mathematical reasoning consists in constructing a diagram” [CP 1.54]. Indeed:

All necessary reasoning without exception is diagrammatic. That is, we construct an icon of our hypothetical state of things and proceed to observe it. This observation leads us to suspect that something is true, which we may or may not be able to formulate with precision, and we proceed to inquire whether it is true or not. [CP 5.162]

In other words, Peirce asserts that “icons”<sup>6</sup> are specially requisite for reasoning” [CP 4.531], while “a diagram [...] is predominantly an icon” [CP 4.418] which resembles the factual relations of its object's parts [cf. CP 2.282]. Therefore, according to the philosopher, the “two great branches of reasoning” are “reasoning by diagrams and reasoning by experiments” [CP 4.74], and he goes further in stating that “the advantage of diagrams in general [is that] one can make exact experiments upon uniform diagrams [which may] take the place of the experiments upon real things” [CP 4.530].

All things considered, as shown at the *top* of the diagram on p. 2, it was thus determined that the empirical methods most consistent with Peircean thought would be the computational generation

<sup>5</sup> See [1.5 – Articulation of the General Theory to the Case Studied](#) for more details.

<sup>6</sup> “Icon” refers here to a *type of sign* within Peirce's typology, rather than any common contemporary use of the term. This will be explained further in sub-section [2.1 \(a\) – Peircean semiotic](#).

of diagrammatic networks of signs, statistically representing the general purpose of selected instances of semiosis. Given my limited resources, this translated into the quantitative content analysis – using the diagram-focused software *KH Coder* – of collective conversations having taken place within select channels on *Internet Relay Chat* (IRC), the latter being, as will be explained later, a somewhat less-studied, predominantly collective conversation platform with which I was extensively familiar.

Moreover, although the conceptual framework put to use here hinges first and foremost on Peircean semiotic, I found it crucial not to limit myself to this theory, as mentioned at the *bottom right* of the diagram on p. 2. Firstly and most importantly, given how one of the ancillary aims of this exploratory study is to highlight the far-ranging applicability of Peirce's theory of signs in communication research, and social science more broadly, I deemed it important to show how theories from diverse major authors can be viewed as fundamentally consistent (and indeed, at times, synergistic) with it. Secondly, as I will seek to demonstrate to some extent, Peircean semiotic – despite, perhaps, the thinker's own hopes in that respect – is not quite a “theory of everything” regarding signification and cognition: it arguably suffers from a number of shortcomings which can be addressed by recourse to other theoretical perspectives.

Consequently, Peirce's thought, including such central tenets as his conception of *semiosis* as a logical and teleological sign-generation process, and that of the mind as an evolving *representation* (defined as *Thirdness*<sup>7</sup>), were supplemented by some of Ludwig Wittgenstein's [1953] key ideas, such as his concepts of *language-game*, *grammar*, *family resemblance*, and *forms of life* to describe

7 This key notion, which is related to the phenomenology developed by Peirce – the categories with which he analyzed any and every experience – will be expounded later on in 2. 1 (a) – Peircean semiotic.

language and the concrete circumstances in which it occurs<sup>8</sup>, in addition to Erving Goffman's [1959] micro-sociological account of every interpersonal interaction as a *performance* (in the sense of a theatrical play, involving a *scene* and *sign-equipment*) and of the self as a *performed character*<sup>9</sup>. Furthermore, my research involves an effort at revealing common features between the theories of Peirce and those of Wittgenstein and Goffman respectively, as well as fundamental ideas shared by all three thinkers.

The interpretative framework resulting from my synthesis of Peirce, Wittgenstein, and Goffman will thus be the foundation for my final, and most general, analysis of the results generated by the methods chosen – this last inquiry being accomplished after a preliminary analysis making use of my familiarity with the topic and the history of the channels studied to identify the probable conventional meaning (or use) of the signs involved.

To summarize, the paramount aim of this work is to make an exploratory case for the adoption, in the scientific study of communication, of a particular kind of theoretical frame of reference (one based on<sup>10</sup>, but not limited to, a Peircean philosophical framework), meant to serve both to justify the use of a certain kind of methodology (diagrammatic and computational) and to interpret the results thereof. Although I do not claim to obtain results that are, strictly speaking,

8 All of which will be described further in 2. 2 – Integrated Theoretical Framework: Peirce, Wittgenstein, Goffman, and are explained in much greater detail in 8. 7 – Appendix 7: Ludwig Wittgenstein: *Philosophical Investigations*.

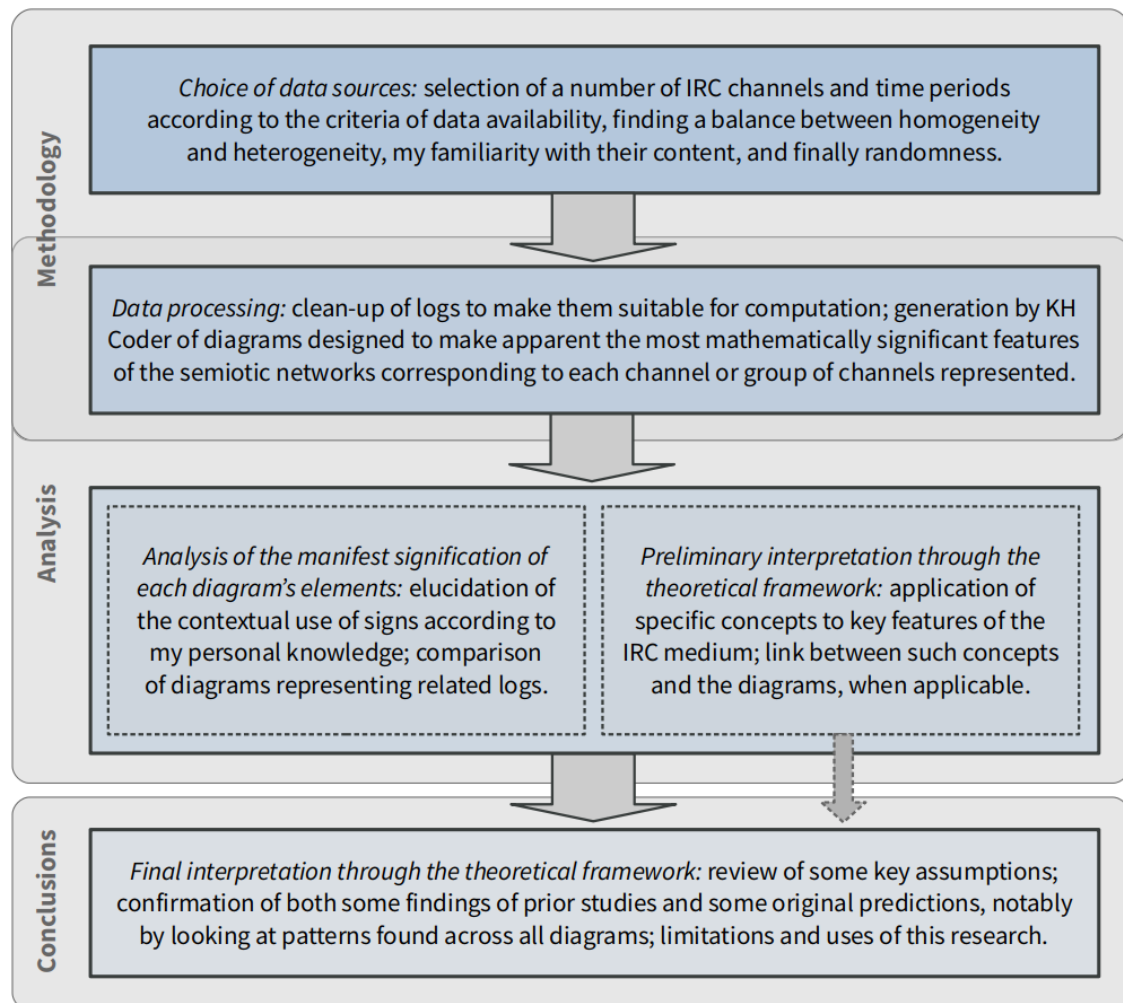
9 Same as the previous footnote #8, except that the relevant appendix is now 8. 8 – Appendix 8: Erving Goffman: *The Presentation of Self in Everyday Life*.

10 Accordingly, Peircean philosophy may herein be considered as the underpinning of a proposed scientific *paradigm* or *research programme*, as will be made clear in section 3. 1 – Epistemological and Methodological Framework (see especially footnote #64). This paradigm or research programme was later found to be closely linked with its proposed methodological counterpart termed *computational semiotic*, introduced in footnote #66 of section 3. 2 – Tools Used and later expanded upon in the sub-section 5. 2 (a) – Each form of semiosis has by-products which are isomorphic to a corresponding type of semiotic network.

statistically generalizable, I hope to demonstrate the soundness of the conceptual framework and methods chosen.

### **1.3 – Case Study: Specific Object, Structure, and Hypotheses**

The empirical case study which constitutes the core of the present work – including the gathering of data along with its subsequent processing, analysis and interpretation – corresponds to the *middle* of the diagram at the beginning of the previous section (on page 2). In turn, I have broken down this part of my thesis into its main components, as shown in the diagram below:



Before I delve into the specifics of this diagram, however, it is worth explaining what *Internet Relay Chat* (IRC) is, and why I chose to study it in the first place. There exists a relatively simple, albeit multifaceted, explanation for the latter: I have spent countless hours discussing with people all over the world via this platform, which nevertheless has arguably been much less widely studied by contemporary social scientists than the likes of Facebook or Twitter, despite yielding text data (so-called “logs”) that readily lends itself to computational analysis. On the whole, this combination of personal experience, relative obscurity, and machine-readable logs is what made IRC appear especially appropriate for my research needs.

As for the nature of this medium, it may be succinctly described as a chiefly *textual* (but also, since the advent of the Web<sup>11</sup>, *hypertextual*) and for the most part *pseudonymous* platform primarily intended for synchronous group conversation – but also allowing for asynchronous and one-to-one conversation, as well as data transfers. Consequently, IRC may be said to combine many of the typical features of Web forums and text messaging. By contrast with ordinary text messaging, IRC puts to use more of the technical features often understood to be those characteristic of “cyberspace”: like Web forums, it allows the user to join or even build online communities (so-called *virtual communities*<sup>12</sup>), which can be briefly described as virtual spaces of shared significations, where people gather by virtue of having similar purposes, interests and beliefs.

11 *Internet Relay Chat* preexisted the *World Wide Web*. According to the relevant *Wikipedia* pages, they were created in 1988 and 1989-1990 respectively. However, *Wikipedia* notes that IRC use has been declining for the past decade, with many users presumably moving to newer, Web-based social media.

12 This expression will hereafter be avoided to highlight the broader philosophical conception of the “virtual” consistent with Peirce’s definition of the word as “A virtual *X* (where *X* is a common noun) is something, not an *X*, which has the efficiency (*virtus*) of an *X*.” [CP 6.372].

Notwithstanding its recent decline, IRC remains the platform of choice for coordinating many large *free and open-source software* (FOSS) projects as well as some non-profit organizations and activist groups, among many others. It is also a popular platform for casual conversation among diverse groups of like-minded people.

In spite of limitations regarding the generalizability of the results of any case study<sup>13</sup>, especially one that is exploratory and does not use a fully randomized sample, there is little doubt to me that the present study of IRC can yield, at least, *some* conclusions that are relevant not only to computer-mediated communication in general, but also to all kinds of communication, by cause of the simple fact that they all share the same general logical structure involving *signs* and *inferences*. It is worth repeating that these conclusions will, however, be limited to providing evidence for the potential scientific fecundity of the theoretical framework and methods chosen.

Let us, at last, move towards explaining the diagram on p. 7, starting with the topmost inner rectangle. Once it had been determined that IRC was to be the medium studied, the first step of my empirical case study proper would have to be the selection of appropriate sources of data. Given how I deemed to have no reasonable hope of yielding *statistically representative* results, this selection would primarily be done according to non-random criteria (as listed in the relevant part of the diagram), designed mainly to facilitate my understanding of the results. Randomness was merely to be the determining factor when other criteria were insufficient to reach a definite choice.

The second step, as described in the second inner rectangle from the top, might be said to serve as a bridge between *methodology* and *analysis*. After obtaining logs from the sources previously

13 Detailed in 3.3 – Sample Selection and Data Collection.

selected, I would have to *process* this data, in two different but interrelated senses: firstly, it would be necessary for me to perform a mostly automated “clean-up” of the logs so as to remove the speaker's pseudonym (and in many cases, the time-stamp) at the beginning of each line, in addition to all so-called “special characters”, and excessively long sequences of characters, among other problems; secondly, the resulting “clean” logs would have to be computed by the *KH Coder* software so as to produce representative diagrams, a process which would surely entail a lengthy trial-and-error process as I experiment with multiple settings among the large number of options.

The third step, described within the third inner rectangle from the top, would involve two closely related aspects of my analysis of the previously generated diagrams. On the one hand, I would have to describe the most likely meaning of their elements – and the meaning of similarities and differences between diagrams representing different data-sets – in accordance with my *personal knowledge* of each channel's *topic* and *history*. On the other hand, I would have to perform, by means of some carefully selected concepts within the chosen theoretical framework, a preliminary analysis of the *attributes* of the IRC medium in general, and of *how these relate* to the diagrams at hand.

The fourth and last step, corresponding to the bottom inner rectangle and to the conclusions of my work, would involve the final application of my chosen theoretical framework to the results of my empirical case study. Apart from a recapitulation of some theoretical assumptions hitherto found to be especially relevant for my research, and potentially others found to be unsuitable, this would subsequently include examining the primary findings validated or invalidated by my case study – notably according to *patterns observed across all diagrams* – and would end with a description of the *major limitations* and the *potential uses* of my research.

Throughout all the aforementioned steps, while the overarching question which initially gave rise to this research endeavour is roughly “Can a Peircean conception of communication<sup>14</sup> be fruitfully used as theoretical foundation for its scientific and empirical study?”, it might already be clear to the reader that, in undertaking to shed light on this matter, I would face a manifold of other riddles. Among them, those that eventually turned out to be the most important – and to which I believe I have found a satisfactory answer, by putting to the test *corollary hypotheses*<sup>15</sup> – include the following (along with, in each case, a possible finding that would contradict and thus falsify my own hypothesis):

- Should empirical scientific research properly based on a Peircean philosophical outlook employ diagrammatic and computational methods, such as the KH Coder software? (*If it should not*, it would prove problematic and inaccurate to understand the diagrams generated by KH Coder according to Peirce’s conceptual apparatus – much like for the question above.)
- Are the theories of other major thinkers compatible with Peircean ones, specifically those of the two authors which I was both well acquainted with and deemed likely to deepen and strengthen Peirce’s semiotic and pragmatism, namely Ludwig Wittgenstein and Erving Goffman? (*If they are not*, in my reading of these authors, I would encounter significant contradictions between their ideas and Peirce’s.)
- Finally, can a case study concerned solely with non-randomly selected IRC logs provide adequate, if only tentative, answers to at least some of the questions above, and others pertaining to semiotic and communication sciences in general? (*If it cannot*, my final interpretation of the results obtained would, at best, yield little else than an exhaustive description of the channels and groups of channels studied.)

#### **1.4 – Overview of the Theoretical Framework**

14 Notably, as consisting in virtual networks of signs and in inferences governed by logical rules and final causation.

15 As ultimately detailed in the conclusions of this work, in section 5.2 – Four Hypotheses Herein Put to the Test and Substantiated by Newfound Evidence.

In choosing Peircean philosophy as foundation for the conceptual framework underlying this project, my aims were twofold. Firstly, I aimed to go beyond, while not rejecting, the popular paradigm of the “social construction of uses” [Proulx, 2005] (or that, virtually interchangeable, of the “social construction of technology”, often advanced as an alternative to “technological determinism”<sup>16</sup>), focusing on digital media viewed as *technical or instrumental devices* embedded in – and determined by – power structures and social institutions. Doubtless, it is a vital aspect of digital media that they constitute *tools*... but my view is that an undue focus on that particular aspect tends to eclipse the most important specificity of such media as unique mechanisms to *share significations*. From the latter point of view, the creative use of new media gives rise to new systems of signs and new uses of existing systems of signs, as stated by French philosopher Gilles Deleuze [1969]:

Meaning is never principle or origin, it is produced. It is not to discover, to restore or to re-employ, it is to be produced by new machineries. [pp. 89—90, translation my own]

Given the exploratory nature of my endeavour, the scientific and philosophical value of the approach suggested – as opposed to established alternatives – remained largely to be seen at the outset. However, I conjectured that my chosen theoretical framework was *complementary* (rather than contradictory in any significant way) with many other research programmes studying the same object, and most importantly, I hypothesized that in spite of inevitable shortcomings, it had the potential to better bring to light aspects of communication that might be often overlooked.

In this line of thought, my second aim was to develop a conceptual framework that could fully account for the *logically orderly*, *purposed* and *predominantly symbolic* nature of collective

16 See notably [Wyatt, 2007].

interactions on IRC, while being consistent with empirical research and epistemological assumptions emphasizing reproducibility and axiomatic transparency<sup>17</sup>. To my knowledge, no research has yet applied Peircean semiotic to the empirical study of online communities, even less done so with recourse to computational and diagrammatic methods, nor with further interpretation through the lenses of author major authors (in the present case, Wittgenstein's philosophy and Goffman's sociology).

The fact that cognition is generally a rule-governed process is strikingly illustrated by meaningful synchronicities such as the simultaneous discovery of calculus by Newton and Leibniz. Similar fortuitous situations occur frequently in cyberspace, such as when two people who do not know each other simultaneously share the same hyperlink at the same place at the same time. However, it occurs to me that we appear to come to a stalemate if we infer from those observations that thought is a mere physical mechanism which can be described through the *very same principles* that govern the material world (a common species of *reductionism*). [Hoopes, 1991, p. 10] In my view, *unique analytical premises* are necessary to account for some properties of *mind* that do not occur as such elsewhere – but can nonetheless be understood logically and explained naturalistically.

The semiotic of Charles S. Peirce, who posited the doctrine of *synechism* to describe the fundamental continuity of everything in nature and culture, but also expected that “a purely psychical way of investigating the mind” [CP 7.504] would arise with advances in scientific knowledge, and “regretted [...] increasingly materialist assumptions and biases” [Hoopes, 1991, p. 13], can offer such premises – and thus a useful middle ground between pure materialism and neo-Platonic or Cartesian

17 See [3.1 – Epistemological and Methodological Framework](#) for further details.

dualism. According to Peirce, the constitutive element of mind is the *sign* (for “[we] have no power of thinking without signs” [CP 5.265]) and its fundamental process, hence underlying sign-generation or semiosis, is that of mediation or representation<sup>18</sup>. The human mind displays a particular kind of semiosis processes, which are otherwise ubiquitous in nature.

In Antiquity, Aristotle identified verbs as signs in *On Interpretation* [~350 B.C.E. (1)], and in the *Rhetoric* [~350 B.C.E. (2)] had already classified signs, according to whether they support “the relation of particular to universal” (such as “The fact that Socrates was wise and just is a sign that the wise are just”) or “the relation of universal to particular” (such as “The fact that he has a fever is a sign that he is ill”) [*Ibid.*]. As Peirce later would, he rightly recognized logical inference as being the essence of a sign's activity. Still, expanding the notion to pertain to *any* state of affairs that *stands for* another only occurred many centuries later: Peirce followed John Locke [1690] in saying that *thoughts are signs*.

Nevertheless, to give my research robust and thorough theoretical grounds, I found that recourse to Peirce's philosophy *alone* was insufficient<sup>19</sup>. In particular, it occurred to me that to demonstrate the fertility of Peirce's theories, it could be of great use to show how they can be integrated harmoniously and complemented with that of other thinkers. Thus, I have added to my final analytic framework, on the one hand, the concepts of Ludwig Wittgenstein, a philosopher noted – like Peirce himself – for ground-breaking advances in formal logic and the study of meaning, and on the other hand, the theories of Erving Goffman, a sociologist celebrated for his contributions to the

18 Corresponding to Peirce's phenomenological ('phaneroscopic') category of “Thirdness” (involving *general laws*), while the categories “Firstness” and “Secondness” correspond respectively to the realms of quality (involving *possibilities*) and of reaction (involving *facts*). These notions will be explained further in 2. 1 (a) – Peircean semiotic.

19 As previously mentioned in 1. 4 – Overview of the Theoretical Framework.

understanding of symbolic interpersonal interactions. I have focused on one particularly well-known and relevant work by each of the two authors: Wittgenstein's *Philosophical Investigations* [1953] and Goffman's *The Presentation of Self in Everyday Life* [1959].

These last two works – and the relevant ideas of their authors, more generally – are each described in an appendix<sup>20</sup>, while the integrated theoretical framework resulting from their overlap and complementarity with Peircean philosophy will be explained in a later section<sup>21</sup>.

### **1. 5 – Articulation of the General Theory to the Case Studied**

Among the most unique features of Peircean semiotic, and one which is unfortunately at times somewhat opaque to comprehension (not least because Peirce's expressed views on the matter evolved with time<sup>22</sup>), is that it ultimately attempts to have explanatory power for the whole of nature, that is, without giving special importance to the human mind in his analysis of the process of semiosis. For Peirce, signs are pervasive in nature, and semiosis is such that it may be said to describe any process involving an object with stands for something to something else. This “to something else” may evidently refer to *somebody*, as Peirce himself often defined signs. However, later in his life, he wrote in a letter to Lady Welby:

My insertion of 'upon a person' is a sop to Cerberus, because I despair of making my own broader conception understood. [SS 80—81]

20 Respectively 8. 7 – Appendix 7: Ludwig Wittgenstein: *Philosophical Investigations* and 8. 8 – Appendix 8: Erving Goffman: *The Presentation of Self in Everyday Life*.

21 Namely 2. 2 – Integrated Theoretical Framework: Peirce, Wittgenstein, Goffman.

22 This appraisal results from a survey of Robert Marty's “76 Definitions of The Sign by C. S. Peirce”, available online on *Arisbe*: <http://www.iupui.edu/~arisbe/rsources/76DEFS/76defs.HTM>. See also the chronological edition produced by the “Peirce Project” at Indiana University, namely, the *Writings of C. S. Peirce*.

Accordingly, as briefly mentioned in the previous section, *purposes* are only one particular form of *final causes*, albeit that “which is most familiar to our experience” [CP 1.211]. The present study, of course, pertains to human communities, and as such deals prominently with human minds and purposes. Although this corresponds to the process of signification-production that the reader is likely to recognize as such most commonly, it is to be understood as a special case of the concept of semiosis developed by Peirce, rather than a perfect or complete illustration of the general principle.

Therefore, I will herewith speak at length of *purposes*, but this is to be taken in the same sense as one can claim that any sign corresponds to a disposition towards a final state of affairs by virtue of whatever has brought it into being. In that sense, whatever the user of an IRC channel *purports to do* is analogous to the regularities apparent in the work of bees, or even in crystals [CP 4.551], and the specificity of human cognition is merely that its determinations *may* rest on conscious volition (in which case we will rightly speak of purposes), although this is *not necessarily* so.

This entails that communication (and cognition) can, in principle, be studied without necessarily referring to human subjectivity. Peircean semiotic can apply to the whole of nature, and human consciousness and purposes are merely a special case of semiosis. Thus, it is possible to study human cognition and communication in much the same way as any other predictable natural phenomenon. We can predict how instances of human semiosis will have regular consequences upon the world, as a result of shared habits of conduct.

## **2 – Theoretical Framework & Literature Review**

### **2.1 – Charles S. Peirce: Sem(e)iotic and Pragmati(ci)sm**

2. 1 (a) – Peircean semiotic:

Semiotic<sup>23</sup>, which nowadays is usually understood broadly as the study of signs and signification, was defined more specifically by Peirce as “the doctrine of the essential nature and fundamental varieties of possible semiosis” [CP 5.488]. For Peirce, in its most comprehensive sense, *logic* itself may be considered synonymous with *general semiotic*, understood as “the science of the necessary laws of thought”, given how “thought always [takes] place by means of signs” [CP 1.444].

This echoes philosopher John Locke’s 1690 *Essay Concerning Humane Understanding*:

[...] [*Semeiotike*], or *the Doctrine of Signs*, the most usual whereof being Words, it is aptly enough termed also [...] *Logick*; the business whereof, is to consider the Nature of Signs, the Mind makes use of for the understanding of Things, or conveying its Knowledge to others. [Locke, 1690, 4.21.4]

Contemporary Peirce scholarship [e.g. Liska, 1996, pp. 16–17; Colapietro, 1989, p. 5] emphasizes the distinction between Peircean semiotic and similar but contradictory theories, usually guilty of what Peirce described as “psychologism”, that is, a reductionist account of the mind explaining all mental processes in one way or another through psychology [See PPM 116]. To take what is probably the most famous rival theory, the “semiology” initially proposed by linguist Ferdinand de Saussure was explicitly considered by its author to be part of social psychology [Colapietro, 1989, p. 50]. By contrast with de Saussure’s endeavour to study the “life of signs in society” [*Ibid.*, p. 5], Peirce’s semiotic consists in an attempt to describe the *fundamental nature of the logical process of signification, that is, sign-activity or semiosis*.

Nevertheless, like de Saussure, Peirce recognized the essentially *social* nature of human sign-

23 Peirce most frequently, but inconsistently, spelled the word “semeiotic”. Following the majority of recent Peirce scholarship, I have preferred the contemporary spelling of “semiotic”.

use: indeed, he emphasized “that all thought is dialogue.” [MS 283: 56 (variant) as cited in Colapietro, 1989, p. xiv] Likewise, he claimed that “[...] the social principle is rooted intrinsically in logic” [CP 5.354] – and we will recall that his account of logic is inseparable from his understanding of semiotic, if not synonymous with it. Still, ostensibly against de Saussure, but in agreement with the later Wittgenstein, Peirce held that “language can only be understood in the concrete context of its uses” [Short, 2007, p. 19], an assumption which is better appreciated in light of his pragmatism<sup>24</sup>.

In the latter sense, one Peircean concept that will be of use for the concrete study of communication in IRC channels is that of *universe of discourse*. Originally a term of formal logic, defined by the *Collins English Dictionary* as “the complete range of objects, events, attributes, relations, ideas, etc, that are expressed, assumed, or implied in a discussion”<sup>25</sup>, Peirce gave the notion a closely related use and meaning as part of his general semiotic (again, held by the philosopher to be in a sense synonymous with logic). In the words of James Jakób Liszka [1996]:

The universe of discourse is what an utterer and interpreter must share in order for communication to result, and so might be more appropriately called a *discourse community*. [p. 92]

Let us now examine the Peircean definition of the *sign*, single constitutive element of mind, thought, and discourse according to Peircean semiotic. Peirce defined the concept in a number of ways proportionate to his interest for it: contemporary semiotician Robert Marty lists no less than 76 definitions of “sign” that he culled from published and unpublished sources.<sup>26</sup> The well-known definition quoted below, of appreciable brevity and intelligibility, was originally published in the

24 See the next section, 2. 1 (b) – Peircean Pragmatism.

25 Online: <http://www.collinsdictionary.com/dictionary/english/universe-of-discourse>

26 See “76 Definitions of The Sign by C. S. Peirce”, available online on *Arisbe*:  
<http://www.iupui.edu/~arisbe/rsources/76DEFS/76defs.HTM>.

*Dictionary of Philosophy and Psychology* edited by James Mark Baldwin:

[A *sign*, or *representamen*<sup>27</sup>, is] anything which determines something else (its *interpretant*) to refer to an object to which itself refers (its *object*) in the same way, the interpretant becoming in turn a sign, and so on *ad infinitum*. [CP 2.303 as cited in Hoopes, 1991, p. 239]

To appropriately grasp the purport of that definition, we ought to highlight the principal distinguishing feature of the Peircean account of the sign: the fact that it is *triadic*, relying on Peirce's categoreal notion of *Thirdness*, introduced in 1885 but initially foreseen in his well-known 1867 essay "On a New List of Categories"<sup>28</sup>. "There is [...] an irreducibly triadic relation among the sign, its object, and the somebody to whom the sign stands for its object." A sign is only a sign insofar as – actually or potentially – it refers to an object and determines an interpretant. [Skagestad, 1998] By contrast, de Saussure's conception of the sign is *dyadic*, its two components being the *signifier* and the *signified*, entertaining an arbitrary relationship between them [See de Saussure, 1916]. Further, de Saussure's "field and sole concern when he [proposed] those notions was to account for the verbal sign, not intending to deal with non-verbal signs (diagrams, facts, etc.)"<sup>29</sup>.

The utmost significance of the concept of Thirdness to understand the mind from a Peircean point of view is encapsulated in the proposition: "[Thirdness] is *Representation* as an element of the Phenomenon." [PPM 167] To understand what is thereby expressed, it is necessary to look into the two other "modes of being" (or "phaneroscopic categories") put forward by Peirce [CP 1.23], namely

27 "A *Sign* is a Representamen with a mental Interpretant. Possibly there may be Representamens that are not Signs. Thus, if a sunflower, in turning towards the sun, becomes by that very act fully capable, without further condition, of reproducing a sunflower which turns in precisely corresponding ways toward the sun, and of doing so with the same reproductive power, the sunflower would become a Representamen of the sun. But *thought* is the chief, if not the only, mode of representation." [CP 2.274]

28 That essay was considered by Peirce to be "the gift [he made] to the world" and his "one contribution to philosophy". According to James Hoopes, "time has vindicated these sentiments" [1991, p. 23].

29 Fernando Andacht, personal communication [2015].

*Firstness* and *Secondness*. Peirce claimed these categories to be phenomenological<sup>30</sup> rather than metaphysical, resulting from “an analysis [...] not [of] what really is, but only what appears to everyone of us in every minute of our lives” [CP 2.84] – that is to say, they are elements of the world of *experience* (or “seemings” [CP 2.197]), and purport to embrace it completely.

Representation being the phenomenological aspect of the category of Thirdness, *reaction* (such as we find in “brute force” [*Ibid.*]) is the phenomenological aspect of the category of Secondness, while that of the category of Firstness consists in *quality of feeling*. [CP 5.63] From a broader perspective, whereas *Firstness* (or “originality” [CP 2.85]) corresponds to the pure abstract possibility found in *qualitative properties*, *Secondness* (or “binarity” [*Ibid.*]) corresponds to the concrete actions or events which make up (dyadic) *factual properties*, and *Thirdness* (or “mediation” [CP 2.86]) corresponds to (triadic) *general properties* characterized by regularity. As implied by the Peircean principle of synechism<sup>31</sup>, the co-existing logical categories belong to a continuum: each of them ceaselessly flows into the others. In addition, because they are *analytical* categories, any aspect of reality can be seen through either of them as a result of the interest of the inquirer [CP 1.23].

The three categories are useful to grasp why Peirce disagreed with theories “[denying] any kind of action except purely mechanical action” [CP 5.63]: mechanical action belongs to the category of Secondness, while Thirdness – no less real than the other two modes of being – is found in “all action governed by reason” [CP 2.86] and thus encompasses *teleological* processes, epitomized by *thought considered as a process of sign-generation or semiosis*. A teleological process is one driven

30 The term “phaneroscopic” was used by Peirce to distinguish his theory from the phenomenology of G. W. F. Hegel.

31 Touched upon in 1. 4 – Overview of the Theoretical Framework.

by *final causation*<sup>32</sup>, a notion traditionally used by Aristotle, who divided “all causation [...] into two grand branches, the efficient, or forceful; and the ideal, or final.” [CP 1.211] These two modes of action are conceivable only in opposition with one another and are “polar contraries” [CP 1.213].

As defined by Peirce, efficient causation is dyadic and concerns *hic et nunc* (“here and now”): “a compulsion determined by the particular condition of things [...] to make that situation begin to change in a perfectly determinate way [independently from] the general character of the result” [CP 1.212]. Final causation, for its part, is triadic and involves potential existence [See CP 1.218], or what Peirce calls “being *in futuro*” (which “appears in mental forms, intentions and expectations” [CP 2.86]): its “mode of bringing facts about” [CP 1.211] amounts to reaching *ends* (that is, “a general description of result” [*Ibid.*]) regardless of the *means* through which this is to be achieved<sup>33</sup>.

Being triadic in nature, the logical structure of final causation is also that of signs, and consequently that of the mind as well, on the grounds that “the mind is a sign developing according to the laws of inference.” [CP 5.313] As Liszka [1996] explains (emphasis mine): “[...] human and mental phenomena are articulated by purpose and intentionality ([a species of] what is traditionally called after Aristotle 'final causes'), *which is the essence of sign activity [...]*” [p. 8]. Furthermore, “[...] final causation is logical causation” [CP 1.250 as cited in *Ibid.*, p. 9]. It follows that the study of signification from the perspective of Peircean semiotic should describe *logical and teleological processes of sign-generation*; through these lenses, an IRC channel is an example of such a process insofar as it consists of a community of people interacting in a predictable fashion which can be

32 Accordingly, sign-generation is teleologically determined by final causation.

33 Peirce’s example: the trajectory of a bullet after it has been fired is a matter of efficient causation, but both the intentional act of firing the gun at a target, and the relation by which a general law causes the action or event of the bullet being fired and following a determinate trajectory, are matters of final (or ideal) causation [CP 1.212].

analyzed and understood in light of the participants' shared purpose.

Peirce made use of his three categories and his triadic definition of the sign to create an elaborate classification system for signs, of which several increasingly complex versions were developed. In 1903, he distinguished three trichotomies resulting in ten classes of signs. By 1907, this had tentatively expanded to ten trichotomies resulting in *sixty-six* classes of signs! At the time of his death, he left unfinished these efforts. For lack of time and space, I will altogether leave out of this document his most sophisticated classification, and omit as well a truly exhaustive explanation of his simpler (and more frequently used by present-day scholars) classification<sup>34</sup>.

To present an abridged rendition of the 1903 classification, a useful starting point is Peirce's very first attempt at sign-classification, part of his 1867 essay on the *categories* [CP 1.558]. This early experiment divided "representations" following the way in which they relate to their *object*: if it is by sharing a quality or trait, they are *likenesses* (later termed *icons*); if it is by accord with a fact or event, they are *indices*; if it is solely by virtue of "there being a habit, disposition, or other effective general rule that it will be so interpreted" [CP 4.447], they are *symbols*<sup>35</sup>. Symbols, being the "objects of the understanding" [CP 1.559], are themselves divided into *terms* (eventually labelled by Peirce *rhemes*, a category encompassing all terms but not limited to them), *propositions* (eventually called *dicisigns*, with propositions being understood as a variety thereof), and *arguments* (also named *delomes*)<sup>36 37</sup>, of

34 Such an explanation can be found notably in Liszka's [1996] *General Introduction to the Semeiotic of Charles S. Peirce* [pp. 34–52].

35 20 years later, in the text "A guess at the riddle", Peirce would develop new terminology according to which the relationships between sign and object defining each of these three categories respectively correspond to a *First*, a *Second*, and a *Third* [MS 909; EP1: 245-279].

36 The remainder of the text of the current section (excluding tables) will use Peirce's later and more distinctive terminology. However, occasionally, an early term and a later synonym will both be included in subsequent sections when deemed appropriate.

37 The *Stanford Encyclopedia of Philosophy* provides as examples: for *rhemes*, "— is a dog", "— is happy", "—

which only the last two varieties are capable of truth or falsehood, and are unique to human cognition.

Arguments may then be either *deductions*, *hypotheses* (later called *abductions*), or *inductions*.

Throughout later classifications, the trichotomy of *icons*, *indices*, and *symbols* would remain the most fundamental division of signs [CP 2.275]. However, in the 1903 classification, Peirce would consider the relation of the sign to each of its components, and not solely to its object, to conceptualize two new trichotomies: the division into *rhemes*, *dicisigns*, and *arguments* was thus reinterpreted as the way the sign relates to its interpretant, while a new division comprised of *qualisigns*, *sinsigns*, and *legisigns* was created to describe the way the sign relates to its “ground”<sup>38</sup> — that is, to itself as an abstract entity. The result was the three trichotomies shown in the table that follows (further subdivisions, omitted here, exist for most categories):

| Three trichotomies of Peirce's 1903 classification of signs |                                             |                                                |                                               |
|-------------------------------------------------------------|---------------------------------------------|------------------------------------------------|-----------------------------------------------|
|                                                             | <i>Firstness</i>                            | <i>Secondness</i>                              | <i>Thirdness</i>                              |
| <i>Relation of the sign to its ground</i>                   | qualisigns<br>(or “tones” /<br>“potisigns”) | sinsigns<br>(or “tokens” /<br>“actisigns”)     | legisigns<br>(or “types” /<br>“famisigns”)    |
| <i>Relation of the sign to its object</i>                   | icons<br>(or “likenesses”)                  | indices                                        | symbols                                       |
| <i>Relation of the sign to its interpretant</i>             | rhemes<br>(or “semes” /<br>“sumisigns”)     | dicisigns<br>(or “phemes” / “dicent<br>signs”) | arguments<br>(or “delomes” /<br>“suadisigns”) |

As a sign necessarily relates to all three of its components, it must include an element from each of the three trichotomies (pertaining respectively to its *ground*, its *object*, and its *interpretant*),

loves —” or “— gives—to —”, etc.; for *dicisigns*, “Fido is a dog”, “Larry is happy”, “Fido loves Larry”, “Larry gives food to Fido”, etc.; for *arguments*, any sign understood as “enabling us to understand the sign as part of a rule governed system of knowledge and signs etc.” [<http://plato.stanford.edu/entries/peirce-semiotics/#Inter>]

38 Peirce defined the “ground” itself as “a pure abstraction” [CP 1.551] (of a quality), or “a sort of idea” [CP 2.228], *in which respect or capacity* a sign stands for an object [*Ibid.*].

although some aspect(s) will always predominate in any given sign. Furthermore, according to Peirce's phenomenology, with respect to the categories corresponding to the *rows* of the table above, a *First* can only determine a *First* (thus, qualisigns are always iconic and rhematic, and icons are always rhematic), while a *Third* can only determine a *Third* (thus, arguments are always symbolic legisigns, and symbols are always legisigns); in addition, with respect to the *columns* of the table above, a category may only be combined with itself or one higher in number. These rules result in ten classes of signs [Liszka, 1996, p. 45] (the use of a grey font indicates redundancy, while the terminology between parentheses tentatively reflects that of Peirce's earlier writings):

- |                                                                |                                                                 |
|----------------------------------------------------------------|-----------------------------------------------------------------|
| 1. rhematic—iconic—qualisigns<br>(term—likeness—tones)         | 6. rhematic—indexical—legisigns<br>(term—indexical—types)       |
| 2. rhematic—iconic—sinsigns<br>(term—likeness—tokens)          | 7. dicent—indexical—legisigns<br>(proposition—indexical—types)  |
| 3. rhematic—indexical—sinsigns<br>(term—indexical—tokens)      | 8. rhematic—symbolic—legisigns<br>(term—symbolic—types)         |
| 4. dicent—indexical—sinsigns<br>(proposition—indexical—tokens) | 9. dicent—symbolic—legisigns<br>(proposition—symbolic—types)    |
| 5. rhematic—iconic—legisigns<br>(term—likeness—types)          | 10. argumentative—symbolic—legisigns<br>(delome—symbolic—types) |

Rather than being strictly separate, these classes generally ought to be viewed as “the manner in which a sign can be abstracted, emphasized, or perspectivized for certain purposes” [*Ibid.*, p. 47].

Many classes are always co-present in each actual sign<sup>39</sup>: whenever they come into being, all signs

39 Peirce explains: “Just as a photograph is an index having an icon incorporated into it, that is, excited in the mind by its force, so a symbol may have an icon or an index incorporated into it, that is, the active law that it is may require its interpretation to involve the calling up of an image, or a composite photograph of many images of past experiences, as ordinary common nouns and verbs do; or it may require its interpretation to refer to the actual surrounding circumstances of the occasion of its embodiment, like such words as that, this, I, you, which, here, now, yonder, etc. Or it may be pure symbol, neither iconic nor indicative, like the words and, or, of, etc.” [CP 4.447] The last sentence in this excerpt may superficially appear contradictory if one does not properly distinguish between a legisign (such as a symbol) *in itself* and a legisign that has been instantiated as a replica.

must be instantiated as sinsigns (“[acting] primarily through [their] singularity, [their] temporality, or [their] unique location” [*Ibid.*, p. 36]), such as when the general idea, law or habit corresponding to a legisign is put into practice. Peirce calls the resulting embodiment a *replica* [CP 2.246 in *Ibid.*, p. 46], although the concept does not necessarily refer to actual, individual sinsigns: symbols can also be replicas of other symbols (for example, two words that mean the same thing in different languages are considered replicas of their shared meaning).

In addition, with each aspect of an instantiated sign, classes corresponding to higher-level categories include classes from lower categories (for example, sinsigns always include qualisigns [CP 2.245 in *Ibid.*], and indices always involve icons [CP 2.248 in *Ibid.*]). As a consequence, the class to which a sign is *primarily* said to belong depends in an essential part on the purpose of the inquiry.

## 2. 1 (b) – Peircean pragmatism:

Pragmatism was defined by Peirce as an elementary normative principle (or “maxim”) of logic [CP 5.402], according to which the *meaning* of a conception – “that is, the rational purport of a word or other expression” [CP 5.412] – is to be found wholly in the *practical consequences* thereof. The term and theory of pragmatism that Peirce created gave rise to a major philosophical movement<sup>40</sup>, due in no small part to the noteworthy influence of two of Peirce's contemporaries: his friend William James and his one-time student John Dewey. James famously made pragmatism into a complete philosophical system and way of life, contrary to Peirce who wished emphasize the precisely logical

Thus, Liszka correctly states that “there is no such *thing* as a pure symbol” [1996, p. 46] (emphasis mine), since “[a] word is not a thing” [CP 4.447], although a legisign and its multiple replicas are reciprocally necessary for either to be meaningful.

40 To this day, it is the only major philosophical movement having originated in the New World.

(and non-psychological) nature of his *pragmatic maxim*. [PPM 8–9; 33]

Peirce was by and large dissatisfied with other interpretations of the theory<sup>41</sup>. He viewed the act of defining a word as an important normative process, especially for the purposes of scientific inquiry – he argued that novel scientific conceptions ought to be safeguarded. On one occasion, he sought to bring together various views of pragmatism by reviewing other definitions to expand and improve his own. [CP 5.494] However, especially later in his life, he was rather inclined to highlight points of contention between his own definition and some major alternatives: chief among them the pragmatism of James, which Peirce found at fault of psychologism, and to erroneously claim that the end of human life was action, while in Peirce's view the only good worthy of being pursued for its own sake was the “development of concrete reasonableness” [PPM 25].

The pragmatic maxim is perhaps most useful for the systematic study of complex concepts. According to the *Internet Encyclopedia of Philosophy*: “Put in its simplest terms, the pragmatic maxim allows us to see what difference the truth of certain concepts would make to our lives.” This can be better illustrated by another of Peirce's formulations of the pragmatist conception of meaning, emphasizing that a word is (only) meaningful insofar as it alters the conduct of whoever uses and believes it<sup>42</sup>:

[...] the meaning of a word really lies in the way in which it might, in a proper position in a proposition believed, tend to mould the conduct of a person into conformity to that to which it is itself moulded.

41 By the end of his life, Peirce rather spoke of “pragmaticism” to distinguish his theory from others that adopted the now famous “pragmatism” label. For the sake of simplicity and readability, I have preferred the more common spelling of “pragmatism”.

42 Peirce equated “the question of pragmatism [with that] of the logic of abduction [or hypothesis]” [PPM 249], for if we admit his definition of pragmatism, it alone suffices to sort between hypotheses that are and are not admissible. A hypothesis may only be rightly considered one insofar as it lends itself to experimental (practical) verification, and two hypotheses are identical if they have the same bearing upon one's practical conduct.

[CP 1.343]

In other words, “the pragmaticist theory of the human subject [emphasizes] habits of action [and] especially [...] those habits that *result from* processes of semiosis.” [Colapietro, 1989, p. 46] According to Peirce, the “sole motive, idea, and function [of thought] is to produce belief” [CP 5.396 *in* Hoopes, 1901, p. 165], while “the essence of belief is the establishment of a habit; and different beliefs are distinguished by the different modes of action to which they give rise.” [CP 5.398] Conversely, “a deliberate, or self-controlled, habit is precisely a belief.” [CP 5.480]

A pragmatic approach to the study of thought (and, it follows, communication, as Peirce held that thought is in essence a dialogue) must therefore take as starting point the premise that “[...] the most perfect account of a concept [...] will consist in a description of the habit which [it] is calculated to produce.” [CP 5.491] The nature of a habit being that of a general rule for action [CP 5.396], this description must include three elements: “the kind of action to which [the habit] gives rise, [...] the specification of the conditions and [...] the motive” [*Ibid.*], that is to say, a habit will be understood as *a disposition to act in a determinate way, under some definite circumstances, for a given motive*. The motive, “an aim or end appearing ultimate to the agent” [CP 1.585], is here effectively synonymous with purpose; the notion of purpose is central to a pragmatic account of meaning, from this perspective taken “to denote the intended interpretant of a symbol” [PPM 232]. Indeed:

[...] the indeterminacy of a sign [is] a latitude [of interpretation] which might affect the achievement of a purpose. For two signs whose meanings are *for all possible purposes* equivalent are absolutely equivalent. [CP 5.448] (emphasis mine)

In addition to the meaning of a concept depending upon a habit of conduct or rule for action,

if the triadic relation constitutive of the sign “is not of a degenerate<sup>43</sup> species, [it] is related to its object only in consequence of a mental association” [CP 3.360]. Such mental connections that relate signs to their objects are polymorphic. “Habits have grades of strength varying from complete dissociation to inseparable association.” [CP 5.477] Although this conclusion applies solely to *genuine symbols* that result from convention and enable logical operations, such symbols make up the fabric of natural human language and therefore the bulk of what the present endeavour aims to study.

## 2. 1 (c) – Semiotic & pragmatism as interdependent frames of reference:

The two primary constituents of the theoretical framework herewith put to use, namely semiotic and pragmatism, were jointly developed by Charles S. Peirce as interdependent parts of his thought. Indeed, according to Vinicius Romanini [2009], Peirce developed his theory of semiotic in no small part “[as an] attempt at offering a strictly logical proof – or, at least, a philosophically consistent one – of his version of pragmatism.” Logic and general semiotic being one and the same for Peirce, it is no surprise that pragmatism, conceived as a proposition of logic, would be regarded by the philosopher as germane to semiotic.

Keeping both theoretical models in mind at once, I am led to interpret in a new light the semiotic notion that “every cognition is determined logically by previous cognitions” [CP 5.265 in Hoopes, 1991, p. 57], that is to say, to interpret pragmatically the logical process of human semiosis as one that is *fundamentally* determined by purpose. Indeed, the logical process of semiosis is, in this view, one of *bringing about ends* – for us humans, this occurs in conformity with our inborn

43 According to Peirce: “Signs have two degrees of Degeneracy”, which correspond respectively to *indices* (degenerate “in the lesser degree”) and *icons* (degenerate “in the greater degree”). [CP 2.92]

dispositions and our acquired habits [see CP 5.487]. Correspondingly, a fundamental area of agreement between Peircean semiotic and pragmatism is metaphysical *realism* and symmetrical opposition to *nominalism*<sup>44</sup>. For Peirce, “what the word *signifies* [...] is real [...] as a description of habit, disposition, or behavior” [CP 1.27 fn P1]. In other words: whether uttered or believed, a conception – in compliance with the habit of conduct corresponding to it – has real consequences upon the world, which display some level of regularity. This, again, refers to the idea of cognition and communication – here, conversations on IRC – as processes logically determined by one’s presence within a specific setting (semiotically speaking, a *universe of discourse*) coextensive with particular conventions and shared purposes.

Peirce’s assertion that “a person is nothing but a symbol involving a general idea” [CP 6.270] can be better understood in light of the notion that a general sign or symbol is a mental connection pertaining to a habit of conduct. Neither aspect is immutable: “every symbol is a living thing, in a very strict sense [...]. The body of the symbol changes slowly, but its meaning inevitably grows, incorporates new elements and throws off old ones.” [PPM 24–25] This conception, which holds that “every general idea has the unified living feeling of a person” [CP 6.270], makes plain why one’s identity develops by the same means as does the meaning (or effective definition) of a general sign. In that sense, the process of cultivating one’s identity amounts to one of *self-definition*, and the result of that process – one’s conception of oneself – has regular yet dynamic consequences upon reality.

To summarize: from the angle of Peirce’s semiotic, what takes place in an IRC channel ought

44 Other conceptual oppositions exist in philosophy involving the concept of “realism”, such as that between “realism” and “anti-realism”. See [footnote #181](#) in [8. 7 – Appendix 7: Ludwig Wittgenstein: \*Philosophical Investigations\*](#).

to be considered as a variety of semiosis; from the complementary angle of his form of pragmatism, the study of IRC must look to the regular practical consequences of what occurs on the platform to understand the meaning of what is communicated there. Understanding such instances of semiosis, or sign-generation, to consist in a process of *bringing about the ends* determined by the purpose of the participants, my aim as researcher should be to describe the *habits of action* (in this case, essentially with regard to the use of language) which correspond to each given purpose, as well as the common “products of sense” or *universe of discourse* serving as precondition for the discussion.

## **2. 2 – Integrated Theoretical Framework: Peirce, Wittgenstein, Goffman**

In the previous section, I emphasized solely the Peircean theories of semiotic and pragmatism. However, among the stated aims of the present endeavour was to demonstrate how Peircean philosophy can be integrated harmoniously with the ideas of other thinkers<sup>45</sup>. Consequently, I decided to supplement my core conceptual framework with the philosophy of Ludwig Wittgenstein and the sociology of Erving Goffman. These particular authors were chosen in part because I was already familiar with them, and in part – and most importantly – because I found their thought to bear striking similarities with that of Peirce, in some respects at least.

My ultimate aim was to devise a theoretical framework that would integrate the thought of the three thinkers into a coherent whole, at once emphasizing common ground in their respective thought and mitigating the gaps in one thinker's theories with the theories of the others, for my

45 As explained in sections 1. 2 – General Object of Inquiry and Aim of this Research and 1. 4 – Overview of the Theoretical Framework.

present research purposes. If required to better understand the present section, the relevant appendixes<sup>46</sup> may be consulted with respect to concepts developed by Wittgenstein and Goffman, respectively.

To begin with, let us examine the features common to the ideas of Peirce and Wittgenstein, as well as potentially complementary concepts. Perhaps above all for the purposes of this work, both agreed with the notion that human cognition and language are not different *in substance* from the rest of nature: as part of nature, they can be studied and understood like any other part of it through phenomenology (Peirce's "phaneroscopy") and logical analysis. Yet, the conceptual apparatus used by the two thinkers appears quite different: oversimplifying, the former describes cognition and communication – as explained in detail in previous sections – as generally purposeful manifestations of sign-activity ("semiosis"), the latter as peculiar kinds of games ("language-games") governed by *grammar* and expressing countless *forms of life* bound by *family resemblance*.

Regardless of these largely superficially different conceptions, however, both philosophers also agreed on the logical structure of words – considered as symbols in the Peircean sense: abstractions, generals, rules or tendencies rather than their actual instantiation – being that of a system of rules. More peculiarly, they both championed the view that *the consequences of the use of words are where their meaning is to be found*<sup>47</sup>, in spite of which they also both depart from behaviourism: Peirce by conceiving signs as teleological processes of representation and acknowledging the intentional nature of the human mind; Wittgenstein by explicitly rejecting the label in refusing to deny

46 8.7 – Appendix 7: Ludwig Wittgenstein: *Philosophical Investigations* and 8.8 – Appendix 8: Erving Goffman: *The Presentation of Self in Everyday Life*.

47 Their conception of the matter is not rigorously identical: while Peirce's pragmatic maxim extended this hypothesis to all signs, Wittgenstein specified (in §43 of the *Investigations*) that not all cases where we employ the word "meaning" correspond to his account of meaning as use.

mental states (in §307–308 of the *Investigations*) [1958/2009, p. 109]. Furthermore, while Peirce believed *semiosis* to be autonomous [Ransdell, 1992, p. 3], the later Wittgenstein agreed (with respect to humans) in asserting that *grammar* is autonomous [Johnston, 2007, p. 389].

Accordingly, considering these additional fundamental similarities, I would argue that one ought to view the different ways the two philosophers conceptualize communication as complementary perspectives. While Peirce wholly agrees with Wittgenstein that conventional systems of rules largely govern human purposeful activity, the latter's notion of *grammar* conceptualizes within a single term, and defines, what such sets of rules are. This same concept of grammar is defined in such a way as to highlight how conventional rules are always embedded into particular social practices – what he calls *forms of life* – a fact which Peirce only tangentially touches upon. Finally, and perhaps most importantly, Wittgenstein's understanding of social practices involves describing the formal relationship between different forms of life as being one of *family resemblance*, a concept which has no direct equivalent in Peircean philosophy – but doesn't contradict it in any way.

As for integrating Peircean thought with that of Goffman, the most valuable common ground to be found for the purposes of this work may concern their respective accounts of the self. As we will recall, Peirce argues that “a person is nothing but a symbol” [CP 6.270]. Goffman, for his part, holds that the “self, [...] as a performed character, is not an organic thing that has a specific location, [but] a dramatic effect arising diffusely from a scene that is presented” [Goffman, 1959, p. 252–253].

They both emphasize the importance of self-representation in the nature of identity and communication – collective or individual. Goffman's [1959] account of social interaction, while the author speaks of the *presentation* of self, could be said to describe a process of the *representation* of

oneself to oneself and to others (understood as the unavoidable functioning of society): what is “presented” does not occur immediately to the performers, but only exists through the representation they have of the “definition of the situation” [Goffman, 1959, p. 106]. The performance can indeed only occur if whoever participates in it has in mind an expectation, which is in itself a representation of what is to take place. Therefore, for both thinkers, it can be said that *the nature of the self is that of a representation*: its meaning is that of ever-changing symbols resulting from a process through which one's place in the world is interpreted by oneself and by the others with whom one interacts.

Despite these numerous key similarities between the two authors' theories, they also feature numerous key differences – but again, such theoretical discrepancies may be understood in light of some crucial aspects in which the two thinkers' ideas are complementary rather than contradictory. For instance, while Goffman's concept of scene may be said to describe, from a Peircean point of view, the preconditions for human purposeful semiosis to occur, and in that sense be somewhat synonymous with Peirce's concept of universe of discourse, Goffman's notion of scene also provides a particular way of conceptualizing such circumstances according to their concrete requirements, starting with what Goffman calls expressive equipment (or sign-equipment), namely the system of interdependent symbols a performer can and/or must use to be part of a given scene. This concept is closely related to the idea of expressive coherence, which describes how a “performer” (or participant) will find that only “a small choice of lines will be open [...] and a small choice of faces will be waiting [...]” [1967, p. 7] when participating in a given situation. Similarly, Goffman's notion of front refers to all the conventional expressive equipment put to use – intentionally or not – by an individual during a performance [1959, p. 22].

All these concepts developed by Goffman might be considered, from one angle, to be subsumed by Peirce's notion of *universe of discourse*. However, Goffman's perspective provides a conceptual framework to describe the various components of the universe of discourse, in particular with respect to how they relate to concrete social situations. People's socially constructed expectations and purposes – which express conventions – essentially determine their behaviour. Meaning is constructed according to a complex interaction between one's own expectations and purposes, and those of others with whom one interacts purposefully. That is, Goffman's ideas help to understand human semiosis in a way that fully incorporates the crucial importance of social relationships.

Some commonalities in the chosen works of Wittgenstein and Goffman were deemed especially relevant to bridge the gaps between the three authors. Thus, the word *play* has in English a double-meaning which is of great relevance to the ideas of both Wittgenstein and Goffman: for instance, IRC interactions display features of both *play* in the sense of Ludwig Wittgenstein's "language-games", and of *a play* in the theatrical sense, as in Erving Goffman's account of "impression management". It follows that participants, as they interact with one another, at once follow *rules* and *roles* – the latter being an upshot of the former when they are carried out or embodied within individuals. The twin analogies of language as a game of chess (Wittgenstein) and of social interaction as a theatrical performance (Goffman) are not perfect nor do they profess to be, but each of them illuminates aspects of communication which the other does not adequately account for. Nevertheless, they have much in common: while social interaction encompasses more than language-games in a strict sense, it *is* also "purposed activity which takes place through signs".<sup>48</sup>

48 My own "working definition" of a "game", as described in 8. 7 – Appendix 7: Ludwig Wittgenstein: *Philosophical Investigations*.

All three authors agree that a characteristically – albeit not quite uniquely – human class of signs (which Peirce called *symbols*), defined by conventional rules, make up the very fabric of human life, thought and language. The normal and inescapable way for humans to be in relation with one another – and further, to relate to themselves – is characterized by habit and rule-like behaviour. They also all assert that the use of signs by humans, including for the most solitary cogitations, has an irrevocably social character. Strictly speaking, nobody ever thinks *alone*; a more accurate way to say that we are “in thought” would be to say that we *take part in the collective process of thinking*. Likewise, “we are in the process of becoming who we want to become (or can't help becoming) as we plunge into that dialogical development of signs called conversation.”<sup>49</sup>

Hence, taking part in an IRC conversation entails taking part in a collective cogitation in the most rigorous and obvious sense. That cogitation can be mapped out (or modelled). The strictly collective nature of the cogitation gives some peculiar characteristics to the maps (or models) thus created. In Peircean terms, such maps are *iconic signs* of the semiosis processes within a given “universe of discourse”. In Wittgensteinian terms, we may call them *pictures* (albeit only sketches) of the “language-games” that took place in the spaces studied. Finally, in accordance with Goffman's theory, they can be construed as an account of the predominant expressive equipment brought into play to foster a particular “definition of the situation”.

### **2. 3 – Review of Peircean Assumptions on Cognition & Communication**

At the core of the theoretical assumptions that give shape to this work, one definition emerges

49 Fernando Andacht, personal communication [2015].

as being of the most critical importance: that of the twin concepts of *cognition* and *communication*, understood as one and the same within the interpretative framework put forward in the previous sections of this work. These concepts may be rightly viewed as more or less distinct or identical, depending on one's purpose and perspective. With regard to the purposes of this thesis and the theories chosen to fulfill them, in keeping with the relevant tenets of Peircean semiotic, the two notions are deemed so deeply interconnected and interdependent as to be indistinguishable at times<sup>50</sup>, in light of which the very same definition may refer to both.

The present endeavour, therefore, endorses an interpretation of cognition/communication as one category among the manifold instances of a much broader phenomenon: *semiosis*. In turn, semiosis is understood as the predictable generation of signs – and incidentally, of signification – according to probabilistic logical rules that share the underlying form of representation (that is, one thing “standing for” another). These rules determine endless possible future inferences, among which some are actualized, contingent upon the pragmatic constraints of every situation.

The inferential mechanism of semiosis can be conceptualized in a way designed to visualize its fundamental structure as a *network of interlinked signs*. Such abstract relationships between signs arise when the rules of probabilistic inference apply to a given state of affairs in order to make it progress towards a given final state (the “final cause” of that logical process). The system used by a mind to commit signs to memory, and subsequently retrieve them, mirrors the general structure of all semiosis processes. Accordingly, all signs are encoded in the brain within *networks of potential*

50 When using the proposed Peircean definition, as opposed to any other definition that might be brought into play, the remainder of this section, and the conclusions of this thesis, will subsequently use the expression “cognition/communication” in lieu of “cognition and communication”, to symbolize the two words being united within a single concept.

*associations*, regardless of which aspect of life is under consideration.

In the case of the uniquely human phenomenon of cognition/communication, it is rules of *meaning* – rules of inference distinctively learned throughout a lifetime of conscious experience and crystallized as habits of thought and conduct – that underlie the logical links between the signs thus generated, and it is *purpose*<sup>51</sup> that corresponds to a final state of affairs that the consequences of sign-generation are pragmatically expected to bring about.

Being a process, semiosis itself is inexorably impermanent and intangible, despite having palpable and lasting consequences upon the world. Correspondingly, if signification is understood as emerging from the networks of inferences continuously engendered in semiosis processes, it ought to be apprehended as inherently *virtual*. That is to say, rather than actually residing in some extant elements that one could find in a definite place and time, signification exists instead in the potential ways in which the signs that compose semiosis may subsequently be connected to other signs (by means of future inferences). In that sense, the elementary dynamic structure of all types of semiosis can be accurately described as “virtual networks of signs”.

The latter description readily applies to cognition/communication. However, the characteristic features of this peculiar species of semiosis – namely meaning and purpose – also imply that whenever we refer to a “sign” when discussing semiosis in general, it is more appropriate to refer to a “thought” (or an “idea”, used as a synonym in this context) when specifically discussing cognition/communication. In accordance with the terminology of Peircean semiotic, a thought or idea

51 The present author's view of complex animal behaviour is that it is ultimately driven by affective or “visceral” determinations, which might perhaps be called impulses, and are not believed to be instances of genuine purpose (tentatively defined as impulses that have been conceptualized, reflected upon and developed into conscious acts of volition).

is a particular species of sign in a way analogous to how cognition/communication is a particular species of semiosis. Similarly, meaning may be seen as a particular *species of signification*.

To rephrase and encapsulate the basic assumptions that are most relevant to this work, the human mind has knowledge of the world only through signs that represent it, which acquire their ever-changing meaning through inferences that result from each person's cognitive/communicative habits, themselves the upshot of a complex interaction between innate and learned dispositions. In the brain, such habits are encoded as *virtual, dynamic systems of interconnected thoughts*.

This encoding process subdivides the brain's informational content into smaller functional units that group together “related” fragments of information – and therefore potential thoughts – with a greater *probability of future mutual associations*. In other words, according to mainstream theories in contemporary psychology:

“The mind forms links between [...] constructs that tend to get activated together, and thus, activating one construct that is part of a network should activate the other constructs of the network. Because information is dispersed in people's minds through networks, people who are 'primed' or exposed to one part of an information network should be more likely to think in other ways that are part of that same network.” [Heine, S. J., 2008/2012, p. 410]

Experimental psychology has, indeed, recently provided further robust evidence for the model of “virtual thought-networks” in accounts of the “priming” effect, defined as the phenomenon where exposure to a given type of stimulus will then temporarily alter in a predictable fashion the interpretation of other types of stimuli<sup>52</sup>, on the grounds of which the stimuli involved (or rather, a

52 One prominent subcategory of the phenomenon is the case of “cultural priming”, describing instances characterized by the initial type of stimulus corresponding to a distinctive aspect of a given culture. For example, an extensive body of research [see Oyserman & Lee, 2008 for a meta-analysis] has focused on the effects of priming with “Individualistic” stimuli (e.g. words such as “I” or “me”) versus “Collectivistic” stimuli (e.g. words such as “we” or “us”).

representation thereof) are inferred to be mutually associated in the mind.

## **2. 4 – Summary and Relevance of Previous Qualitative Research on IRC**

IRC has been studied by numerous researchers, often applying the method of “virtual ethnography” [e.g. Hine, 2000] or closely related ones such as “ethnomethodology” [e.g. Lawson, 2008]. The pioneer of research on IRC communities is doubtless the venerable Howard Rheingold, who studied such communities as early as the end of the 1980s and coined the expression “online community” in his landmark 1993 book – published at the very beginning of the *World Wide Web*, while it was still used nearly exclusively by academics. In *The Online Community*, he describes his personal experience as a participant to IRC communities, doing so in the form of a narration filled with anecdotes and marked with a sense of wonder, a feeling which is not unknown to myself.

Regrettably, perhaps illustrating the frantic pace at which the digital realm evolves, to the eyes of the contemporary observer, his efforts are mainly of historical rather than analytical interest: the ethnographic observations on which he bases his analysis have been made long before the days of the *Web 2.0* and *HTML5*, and in all likelihood do not represent the current state of uses of the platform. His analysis highlights the specificity of IRC as being the reduction of simultaneous conversation to text alone [Rheingold, 1993/2000, p. 181], in a time when the addition of *text messaging* features to cellphones was still in its infancy<sup>53</sup>. Today, Rheingold's description would seem to apply to traditional “text messages” rather than to IRC, as the latter is now nearly always enlivened with various hyperlinks pointing to content that is often multimedia<sup>54</sup>. That said, one unique characteristic of IRC

53 The first “SMS” message is sent in 1992, as Rheingold's study is already nearing its end.

54 Today, the advent of “smartphones” allows IRC users to access their favourite network(s) from their cellphone.

which remains unchanged since Rheingold's days is the typically *collective* nature of conversations – which distinguishes the platform from text messaging.

Despite these shortcomings with regard to his description of IRC specifically, Rheingold suggests many broader paths for analysis that are still relevant to academic research on cyberspace in the current context. He explores for instance the notion of “real-time tribes” [*Ibid.*] – evocative of Marshall McLuhan's [1962] “global village” – and the paradoxical relationship participants to online communities have to *proximity* (intense affective proximity can exist in spite of immense geographical distances), as well as the related question of the *strength of interpersonal links* in cyberspace.

In a chapter added to the new edition [2000], Rheingold questions the use of the expression “online community”, noting that sociologist Barry Wellman prefers the presumably more accurate expression “online social network” [Rheingold, 2000, p. 359]. He also draws attention to the hopes of a democratization of the “public sphere” brought about by near-universal participation in an open and decentralized cyberspace [*Ibid.*, pp. 371—372] – but also, to assuage utopian discourses, the risk of “panopticism”<sup>55</sup>, and the critique of “cultural industries” of the Frankfurt School of Theodor Adorno and Max Horkheimer [*Ibid.*, pp. 372—391]. Indeed, the question of knowing whether digital media essentially only reproduce the dynamics of the existing capitalist system, or offer the possibility of genuinely new forms of social and economic relationships, is still open and hotly debated<sup>56</sup>.

More recently, two fellow French Canadian researchers have studied IRC before myself: Madeleine Pastinelli [1999, 2002, 2006] and Guillaume Latszko-Toth [2010]. The former, who

55 This notion, introduced by philosopher Jeremy Bentham and famously developed by Michel Foucault, is especially contemporary in light of the recent revelations of whistle-blower Edward Snowden on state surveillance in cyberspace.

56 In the same vein, we may mention the problems raised by the “digital divide” [See Rheingold, 2000, p. 377].

describes IRC as a medium that completely ignores frontiers and distances [Pastinelli, 1999, 2002], studies amongst other topics the relationship to space of participants to de-localized communities – curiously enough, by relying solely on the observation of rather peculiar channels frequented nearly exclusively by Quebecer adults [*Ibid.*] – and to the notion of “real-time” applied to synchronous IRC interactions [Pastinelli, 2006]. The latter, who performs an analysis of great breadth reflecting the requirements of a doctoral thesis, speaks of IRC in the following terms:

The problem of an IRC network is that it is completely open and largely decentralized, it is a constellation of servers that share a same database on users and discussion 'channels'. [Latzko-Toth, 2010, p. iv, translated from French]

Latzko-Toth puts forth the hypotheses of a continual transformation of IRC, of an intermittent development of the device (through “socio-technic crises”), and of a determining role of “technical virtuosity” in the “mode of sociability” defined by the platform [*Ibid.*, pp. 29—30]. He concludes by confirming the sum and substance of his hypotheses, underlining like Rheingold the “tribal” aspect of online communities, and the “utopian” nature of IRC networks... in the etymological sense of “non-place” [*Ibid.*, pp. 315—337], in this way agreeing with the analysis made by Pastinelli [2006], who draws inspiration from the distinction made by Marc Augé between the *place* (*lieu*) and the *non-place* (*non-lieu*) [p. 199]. Pastinelli's analysis of an IRC network or channel as a “non-place” does not appear appropriate considering the observations made in this work regarding the nature of the online place circumscribed by IRC<sup>57</sup>, and even less so considering Augé's examples of motorways, hotel rooms, airports, or supermarkets [Cresswell, 2004]<sup>58</sup>.

57 Perhaps most explicitly in 8.8 – *Appendix 8: Erving Goffman: The Presentation of Self in Everyday Life*.

58 Wikipedia defines Marc Augé's notion of *non-place* as “the ambivalent space that has none of the familiar attributes of place - for instance, it incites no sense of belonging” (See: [https://en.wikipedia.org/wiki/Marc\\_Augé](https://en.wikipedia.org/wiki/Marc_Augé)

## **2. 5 – Two Examples of Statistical Analysis Applied to IRC Logs**

Prior to the present research endeavor, at least two Master's theses, written twelve years apart, have made use of statistical methods to study IRC interactions. The first one, written by Wernfrid Doell and published in 2000, consists in a linguistic analysis of several IRC channels. Its main goals were the comparison of global grammatical patterns across channels and with other forms of written or spoken content. Its most important finding, as stated by the author, is the notable similarity between IRC conversations and spontaneous casual spoken discourse.

Being almost two decades old, the study evidently did not make use of elaborate computational techniques. Thus, it is strictly limited by its sole reliance on elementary descriptive statistics such as word frequency distribution and lexical density<sup>59</sup> to draw its conclusions. The following constitutes a short analysis of important empirical limitations faced by the aforementioned thesis, especially in relation to how the present study might overcome some of them.

Crucially, for every statistical comparison made with respect to different elements of the IRC text corpus (versus other written texts or spoken texts, for instance), such as the use of particular pronouns, the author's interpretation remained fuzzy and often required specific examples of sentences from the original data-sets to clarify. Indeed, a meaningful explanation of typically subtle differences in the use of pronouns – and other general-purpose words – within a given form of text is necessarily based, to some extent, on the researcher's knowledge of a channel's subject matter, a

[%C3%A9#Second Stage](#)), which seems rather inadequate to describe the purposed interactions in an IRC channel, which is a *community* defined by a *topic* – a word derived from the Ancient Greek τόπος, “place”!

59 In the thesis referenced in this section [Doell, 2000], “lexical density” is described as a measure of the proportion of unique words on the total number of words within a given data-set.

requirement that Doell [2000] attempted to satisfy by manually providing examples of sentences in which the pronouns typically occurred. The need to include a word in its qualitative context could be made considerably more rigorous by the use of diagrams illustrating the contextual network of word use, as is done in the current work.

Finally, the author mentions that the software used in their study does not allow, for instance, to categorize passive verbs without first identifying all verbs in the corpus as such, which entails the need to rely on estimates to perform an analysis of “syntactical complexity” – for which passive verbs are typically used as indicator. Again, the diagrams produced for the present work allow to visually represent word-associations, which can notably be used to know how a given verb tends to relate (passively or actively) to other words.

The second thesis, by Marvin O. Michels [2012], involved the development and subsequent testing of a computer program intended to facilitate the real-time analysis of conversations on IRC – with the ultimate goal of automatically detecting keywords deemed “suspicious” in various channels. Although that thesis belongs to the field of computer science, it is nonetheless relevant to the present study as an account of the kind of computational methods that have been hitherto applied to textual content on IRC. However, for the purpose of the current thesis, only one of the tool’s components will be further explained: “topic analysis”, with respect to which the author initially discusses the existence of interesting probabilistic algorithms to infer “topics” (namely *Naive Bayes*, *K-Nearest Neighbor*, and *Support Vector Machines*), but in the end implements merely a synonym-based topic analysis method to facilitate live-data acquisition by avoiding “any lengthy calculations”.

Thus, the procedure used to reveal current “topics” initially involves a basic word frequency

calculation to reduce the data-set to only the most used words (minimum frequency being determined by the experimenter), after which the remaining words undergo a stemming procedure, and finally are combined with available synonyms throughout the data-set to form general concepts. The resulting synonym-combination born words recurrently went through additional stemming and following synonym combination operations, until word reduction through combination of synonyms was no longer possible. The final “irreducible” collection of words were retained as the “Current Topic(s)”.

This procedure was applied to two different sets of IRC logs, the first from the #ubuntu channel – also included within the channels analyzed in the present work – and the second from the main channel used by the “hactivist” group *LulzSec*. Unfortunately, severe technical shortcomings in both the stemming and synonym combination algorithms proved to generate highly unsatisfying results with respect to topic analysis.

In fact, *synonym-based* topic analysis seems to be a relatively less effective, and overly rigid, method for uncovering overreaching themes in text data. For instance, when used for “query expansion” – which consists in broadening the scope of terminology used to retrieve information, such as when searching for documents containing specific subject matters (or topics) – a synonym-based method was shown to be out-performed by a *topic model-based* method in terms of relevant information detection [Zeng et al., 2012].

As opposed to synonym-based expansion which relies on a standard database of synonyms and lexical variants (or “thesaurus”), a topic model-based one would include additional terms contingent on their co-occurrence with the original query terms in a given context. That is, in practice, a text mining model is trained with a set of independent, but relevant, documents in order to

uncover clusters of words that statistically co-occur – named “topics” – after which terms that figure in the clusters also containing words from the initial query are added to the query expansion.

While synonym-based methods depend on controlled vocabulary sources, topic model-based methods rely on actual data from relevant texts. Thus, the latter crucially depends on the notion that the co-occurrence of words in a document can be predicted by a given topic.

### **3 – Methodology**

#### **3.1 – Epistemological and Methodological Framework**

An appropriate introduction to the epistemological presuppositions underlying this work appears to be a reflection on the necessary criteria to assess whether or not knowledge can be considered “scientific”<sup>60</sup>. We may recall that for 20<sup>th</sup> century philosopher Karl Popper [1934/1959], scientific knowledge is distinguished from metaphysics by *falsifiability*, namely the possibility of its categorical refutation as the logical consequence of a contradictory observation. Likewise, Peirce espoused epistemological *fallibilism*, “the doctrine that our knowledge is never absolute but always swims, as it were, in a continuum of uncertainty and of indeterminacy” [CP 1.171], which entails that from a scientist’s point of view:

The best that can be done is to supply a hypothesis, not devoid of all likelihood, in the general line of growth of scientific ideas, and capable of being verified or refuted by future observers. [CP 1.7]

In the sense of Popper’s and Peirce’s words, all scientific methods consist at their core in *observation*; indeed, Peirce asserts that “all knowledge whatever comes from observation” [CP 1.238

60 I will herewith avoid the debate as to whether or not such a criterion is necessary, judging that this debate is not relevant for one who assumes the possibility of objective knowledge systematically recognizable as such.

as cited in Liszka, 1996, p. 4]. It is worth noting that the concept of observation is here much broader<sup>61</sup> than its narrow meaning in the context of contemporary scientific methodology, and is closer to the notion of *experience* as used by John Locke – quoted approvingly by Peirce<sup>62</sup> – to say that “all the materials of reason and knowledge [come] from experience” [Locke, 1690 as cited in CP 5.611]. Whether one uses the broader definition or the narrower one, however, observation ought to be considered irremediably fallible. As they have just been described, Peirce’s fallibilism and Popper’s falsifiability are both wholly consistent with the basic epistemological orientation of the present work.

Imre Lakatos [1970], one of Popper’s students, developed the latter’s theories notably through the lens of the work of fellow epistemologist Thomas Kuhn [1962/2012], best known for developing the notions of “paradigm” and “paradigm shift” to account for scientific revolutions. Without contradicting Popper’s theories, Lakatos put forth the idea that scientificity is founded on the *predictive power* of competing *research programmes* composed of fundamental and ancillary propositions. The notion of predictive power can be understood as closely related – if not, in practice, synonymous with – to that of *reproducibility*: a prediction made by a research programme is to be considered true if and only if a corresponding outcome can be consistently reproduced<sup>63</sup>.

While Popper and Lakatos both highlighted the importance of the *scientific method* to attain objectivity, I would argue that the concept of research programme developed by the latter, and to a

61 In fact, this definition of the concept is so wide-ranging that even mathematics can be said to be “observational, in so far as it makes constructions in the imagination according to abstract precepts, and then *observes* these imaginary objects, finding in them relations of parts not specified in the precept of construction.” [CP 1.240] (emphasis mine)

62 Peirce himself defined “experience” as “the cognitions which the history of our lives *forces* upon us” [CP 2.784] (compare with the scientific notion of *evidence*), and elsewhere wrote that “*experience*, from the very essence of the word, consists of our belief about a universe – ‘the truth’ – over against our opinions and beliefs” [CP 8.294].

63 This requires any scientific prediction to include a *decision procedure* to determine whether or not this is the case, preferably with as little ambiguity as possible.

lesser extent, Kuhn's similar concept of paradigm, are specially useful in that they highlight the importance of what might be termed *axiomatic transparency*. Indeed, in this view, which is also my own, it is impossible for a scientist to contribute to the advancement of knowledge without relying on some "*unproven*" *assumptions* – analogous to "axioms" or "premises" in formal logic, and corresponding roughly to what Lakatos describes as "fundamental propositions" in this context – to interpret what is observed or experienced. A scientist's duty is thus to be as keenly aware as possible of important assumptions, and to explicitly state such assumptions within one's work so as to make it possible for others to deliberately follow the same reasoning to consequently reach the same conclusions.

Moreover, although I am generally sympathetic to the methodological framework inherited from the tradition of "exact" sciences, which ostensibly tends to emphasize predictive power or reproducibility above all, I am not satisfied by the "positivistic" epistemological stance which often accompanies it.

Chemist and epistemologist Michael Polanyi [1958/1962] remarks that the scientific process depends irremediably on the personal investment of researchers, who are necessarily part of a complex network of social interactions – the "scientific community" – which inform their contribution to the body of scientific knowledge. "Theoretical research is a form of social practice." [Eco, 1976, p. 29 in Colapietro, 1989, p. 33] This conclusion is, here again, reminiscent of Peirce, for whom a *true conception of reality* presupposes the existence of a *community of inquirers* [Hoopes, 1991, p. 55]. According to both Polanyi and Peirce, scientific knowledge – and, more generally, all *demonstrably true* knowledge – implies *transparent dialogue* and *confirmation by peers* (implying,

again, axiomatic transparency and reproducibility). Polanyi is further echoed by sociologist Norbert Elias [1983/1993], who highlighted that there is a constant tension between engagement and detachment in the course of research for the social scientist. In the same line of thought, Howard Rheingold [1993/2000] notes that the role of the researcher cannot be *only* that of a cold and disengaged observer: community is as much a matter of emotion than of reason and data [p. Xxxi].

As a matter of fact, comprehending complex social realms presupposes the subjective investment of the researcher in two interconnected ways: firstly because the study of a given aspect of reality (social or not!) requires the personal interest of the researcher for that particular aspect; secondly because it is a precious tool of the social scientist to be capable of empathizing with the object of his study, allowing him to grasp in depth the meaning of the behaviour of social actors<sup>64</sup>. It is my view that the social scientist's active participation in the realms studied is not only unavoidable but – at least when done deliberately and explicitly – leads to a better understanding of one's object of inquiry than a futile attempt at eliminating all traces of personal engagement.

To summarize what was previously stated in this section, my approach was primarily guided by the scientificity criteria of *axiomatic transparency* and *reproducibility*<sup>65</sup>. These combined criteria<sup>66</sup> have sought – for the purpose of yielding *factual predictions* (and maybe, eventually, something

64 This second observation is a pillar of Max Weber's well-known “comprehensive sociology”.

65 Itself conceptualized as synonymous with predictive power, and as a more directly operational counterpart to the criterion of *falsifiability* (or *verifiability*).

66 In the current context, a new criterion which could harmonize the notions of reproducibility and axiomatic transparency with the increasing prominence of computational tools in contemporary research, is the *requirement of computability*, suggested notably by Pierre Lévy [2011]. This criterion, which appears wholly compatible with the perspectives of Popper and Lakatos, can be defined as the requirement for a scientific theory to be axiomatized and encoded in such a way that (at least potentially) it can be translated into an algorithm that all computers can execute in essentially the same way. However, this does not imply that it is ever unnecessary to ultimately *interpret* the results of such scientific algorithms.

useful<sup>67</sup>) – to give epistemological priority to the results of exact methods and empirical research, while acknowledging two inevitable obstacles which make doubly impossible a *purely* empirical, “fact-based” (or “data-driven”) study of human communication:

- firstly, for any scientific endeavour, one cannot avoid making use of *theoretical assumptions* (such as those systematized into a paradigm or research programme);
- secondly, for social science in particular, *personal investment* and *empathic interpretation* are necessary (indeed, the very nature of social science makes some level of proximity between researcher and object desirable, and interpretation through the lens of personal experience can be a cornerstone of a social scientist’s work).

An approach which would make use of qualitative methods – involving the characterization of the object of inquiry through semantic and conceptual lens – without recourse to quantitative data, such as is ostensibly found in the works of Rheingold [1993/2000], Pastinelli [1999, 2002, 2006] or Latzko-Toth [2010], tends to produce results that are arguably akin to a “story” (that is, appearing mostly anecdotal and/or biographical) and very unlikely to be *reproducible* (or *computable*) in any straightforward way. Conversely, quantitative methods – involving the attribution of numerical values to represent formal or logical features of the object of inquiry – employed without a sufficient qualitative counterpart, as I contend we find in the studies accomplished by Doell [2000] or Michels [2012], albeit highly rigorous, can hardly ever suffice to shed light on the complex relationships between a discourse, the context of its production, and the context of its reception. Finally, regardless of whether qualitative methods, quantitative methods, or both are used, interpretation according to a chosen theory is required to contemplate the *meaning of facts* beyond what is immediately obvious.

67 As the reader may surmise, it follows from the pursuit of *computability* that this could be some kind of computer program.

Consequently, the predominantly semiotic paradigm upon which this research is predicated involves assumptions which have been spelled out at length in previous sections<sup>68</sup>, and the methods chosen to carry out its empirical component attempt to simultaneously be as reproducible as possible while leaving ample room for explicit personal engagement. This ultimately involved the simultaneous and interdependent recourse to *observation (as personal experience)* and *quantitative (or statistical) content analysis*.

For the purposes of the present study, observation consists in my personal experience as a participant (active or *idle*) in some chosen IRC channels. I have privileged this simpler method over other qualitative methods such as ethnography, interviews or focus groups mainly due to my limited resources, as well as to allow for more flexibility in my approach, but also to avoid ending up with a “story” as defined above. On the one hand, by contrast with statistical analysis, observation usually occurs at the very moment the IRC exchanges studied are taking place to grasp their context, shedding light on the subjective experience and intentions of actors. On the other hand, its results alone are not easily reproducible (or computable) and typically have weak predictive power.

For its part, quantitative content analysis as put into practice herein consists in the computational analysis of the formal structure of IRC logs so as to produce a geometrically accurate diagrammatic representation thereof – the results thus obtained are, however, to be *explained* through a combination of personal experience and theoretical assumptions. I have chosen to analyze pre-existing data over the main competing model for quantitative research, namely analysis by means of questionnaires deliberately constructed and administered (the archetype of which being polls), again

68 Indeed, it might be said that one of the most crucial aims of the present work is to show how Peircean semiotic and pragmatism can be used as foundations for a potentially fertile scientific *paradigm* or *research programme*.

due to limited resources, but also and perhaps mostly as the content I sought to analyze was readily available to me. In short, I saw little reason to look further than IRC itself to understand it.

### **3.2 – Tools Used**

For the quantitative component of this study (which was itself made part of this work to increase the *reproducibility* of its results<sup>69</sup>), several analysis techniques borrowed from *content analysis* or *computational linguistics*<sup>70</sup> were judged most appropriate for working within my preferred epistemological and interpretative frameworks. First and foremost, I hypothesized that such techniques, especially among those available for the scientific study of text-based online communication, might be the most consistent with Peirce's philosophy<sup>71</sup>. Although of lesser importance, it also was my belief that among all relevant options, the choice of *free and open source software* (FOSS) would be more consistent with *axiomatic transparency*<sup>72</sup>, given how the underlying algorithms may be freely perused.

Accordingly, and being a proponent of FOSS myself, the main tool herewith put to use is the powerful text analysis program *KH Coder*<sup>73</sup>, programmed primarily by Prof. Koichi Higuchi from the

69 One of the two primary epistemological axioms which guided my research, as explained previously in 3.1 – Epistemological and Methodological Framework.

70 As explained in section 8.9 (a) – Content analysis & computational linguistics, the methods of statistical analysis employed for the present study, taken in isolation, could potentially be categorized within either one of the well-established paradigms of *computational linguistics* or *content analysis*. However, because this endeavour rests on theoretical grounds (and correspondingly has purposes) which are in some respects far from the ones typical of these paradigms, it may be more appropriate to classify its statistical component within a novel methodological paradigm which might tentatively be called *computational semiotic*.

71 As initially explained in section 1.2 – General Object of Inquiry and Aim of this Research.

72 The second of my two primary epistemological axioms, described in 3.1 – Epistemological and Methodological Framework.

73 As *free and open-source software*, it can be freely downloaded on *Sourceforge* and legally shared with your friends: <https://sourceforge.net/projects/khc/>.

Ritsumeikan University of Kyoto. Otherwise, the *Quassel* IRC client and the *Bash* scripting language<sup>74</sup>, on the Kubuntu operating system (an Ubuntu derivative), were used for the collection of data, chosen ultimately because of my familiarity with them. Finally, the *Perl* programming language – in which KH Coder is programmed – was used for pre-processing data. All the software listed in this paragraph is FOSS.

KH Coder, which was used for the nucleus of the analysis that makes up the present case study, is quantitative content analysis software predominantly featuring diagrammatic tools. It was found to ostensibly satisfy the constraints previously put forward in accordance with Peirce's theory of signs, namely that the method used:

- should generate diagrammatic representations;
- these diagrams should consist in networks of the signs, and the logical links between them, that constitute communication;
- the general idea each diagram conveys should partly or wholly correspond to the prevailing shared purpose of the collective conversation represented.

No other FOSS considered for this research *appeared* as consistent as KH Coder with the aforementioned Peircean notions. Its *actual* consistency with these notions, however, remained in need of further confirmation through putting the software into practice and interpreting its results – and that would still say little about the appropriateness of any other software by the same standards.

### **3.3 – Sample Selection and Data Collection**

A case study has as its main limitation the fact that any sample that is not completely random

<sup>74</sup> Typically used in the *GNU/Linux* command-line interface.

and sufficiently large may not be said to be strictly statistically representative of whatever *population* one purports to study. I have attempted to (partly) work around this limitation in two ways: firstly, and perhaps most crucially, by selecting many different groups of IRC users, and many different time-frames of diverse dates and lengths, to be analyzed separately then compared with one another; secondly, by invariably analyzing all participants of the groups studied.

Still, the problems raised by *generalizing* the data must be stressed. According to some thinkers, the researcher tends to choose cases enabling him to validate his theoretical presuppositions [Becker, 2000 in Latzko-Toth, 2010, p. 94]. Therefore, at the very least, conclusions drawn from the cases studied must be nuanced to account for their particular characteristics. Symmetrically, it is appropriate to suggest further case studies to explore different aspects of the same methods.

Thus, four individual IRC channels and three large groups of IRC channels<sup>75</sup> were analyzed for the purpose of this work: two of the small channels I administrate – allowing for *participant observation* – and many other, often larger, channels in which I have never participated at all but which have to do with topics I am familiar with – yielding *non-participant observation*. More specifically, the data sources fall into either one of two categories:

- 1) Public channels which I administered myself during the period studied, namely #politics and #libertyinmind.
- 2) Public channels associated with well-known FOSS and/or community-driven projects which make publicly available their logs online. This includes all channels related to the Ubuntu operating system and its derivatives, the logs of which are freely available at the URL <http://irclogs.ubuntu.com/>, and (for an appendix<sup>76</sup>) several channels related to the MediaWiki software and the Wikimedia Foundation,

75 Albeit two of these groups are primarily analyzed in an appendix, 8. 10 – Appendix 10: Additional Diagrams – Wikimedia-related channels, and otherwise only tangentially mentioned in the conclusions.

76 That is, 8. 10 – Appendix 10: Additional Diagrams – Wikimedia-related channels.

the logs being freely available at the URL <http://bots.wmflabs.org/~wm-bot/logs/>).

This interplay of participant and non-participant observation is meant to reflect my belief that both proximity and detachment generates valuable results for the social scientist, and that those results tend to be complementary of each other rather than contradictory. Still, even in the case of the channels in which I made a point never to participate, it cannot be said that I have been an “absolutely detached” observer, as I have good knowledge of their subject-matter (Ubuntu and Wikipedia).

The official #ubuntu channel on the Freenode network is an exceptionally large channel, home to about ~1600 users, the vast majority of which are inactive (in IRC parlance, “*idlers*”) at any given point. Meanwhile, #ubuntu-offtopic is much smaller but quite large still, having ~200 users. Both channels, as with nearly all Ubuntu-related channels, are very busy – with very few moments of “silence”. By contrast, the official #wikipedia channel is rather empty despite its ~200 users, and #mediawiki is about as empty as #wikipedia despite having over 250 users. In both channels (not specifically analyzed for the present research), it seems that *the rule* is for the vast majority of users to *idle* at all times, and all users to be *idling* most of the time. Much smaller but more close-knit channels such as #libertyinmind (hovering around 10 to 20 users at the time of this study) can be much more active than channels an order of magnitude larger in terms of number of users connected.

The channels selected were chosen both for *convenience* (I had easy access to, and good understanding of, my own channels and official public channels of widely-known collaborative projects) and for their *analytical interest* (channels were initially selected among others because they seemed especially interesting, and very similar as well as somewhat similar and very dissimilar channels were all chosen to allow for more diverse comparisons). Only after all such criteria were

exhausted did I select channels at random. Further research may attempt proper random sampling of channels for increased statistical validity.

Two means of collecting data (IRC logs) for analysis were used. Firstly, I have logged several channels myself, either fortuitously as I happened to administrate them (in two cases), or specifically for the purposes of this research (in the other cases). Such channels, which I personally recorded and directly observed, are the following (although only the first three were, in the end, specifically analyzed in this thesis):

- #politics on the *Rizon* IRC network: from April 29th, 2012 to August 4th, 2014.
- #libertyinmind on the *Freenode* IRC network: from October 29th, 2013 to August 4th, 2014.
- #ubuntu-offtopic on the *Freenode* IRC network: from August 8th, 2014 to October 13th, 2014.
- #ubuntu on the *Freenode* IRC network: from August 6th, 2014 to October 13th, 2014.
- #wikipedia on the *Freenode* IRC network: from September 16th, 2014 to October 13th, 2014.
- #mediawiki on the *Freenode* IRC network: from September 16th, 2014 to October 13th, 2014.

Secondly, the Ubuntu project makes freely available the logs of all Ubuntu-related channels<sup>77</sup> since July 2004. The *wget* tool on GNU/Linux was used to automatically download all text files while simultaneously excluding all other files and preserving the folder structure:

```
wget -r -l5 -A.txt http://irclogs.ubuntu.com/
```

In the same way, the Wikimedia Foundation makes freely available the logs of many related channels since 2007. I downloaded the entire remote folder, then deleted all irrelevant sub-folders.

```
wget -r -l5 -A.txt http://bots.wmflabs.org/~wm-bot/logs/
```

<sup>77</sup> In October 2014, this included the list which can be found in [8.2 – Appendix 2: List of Publicly Logged IRC Channels Analyzed](#). For comparison purposes, an exhaustive list of Wikipedia-, MediaWiki- and Wikimedia-related channels can be found at the URL <http://en.wikipedia.org/wiki/Wikipedia:IRC>.

Some channels such as #wikimedia-labs-nagios, #wikimedia-tech, #wikimedia-toolserver, #wikimedia-wikidata, #wikipedia-en-GAN, #wikipedia-snuggle, and #wikipedia-teahouse were excluded because they largely consist of text automatically generated by “bots”<sup>78</sup>. Also, channels in a language other than English were excluded for all but one data-set including *all* Ubuntu-related channels<sup>79</sup>.

### **3.4 – Ethical Considerations**

The data analyzed did not raise any meaningful ethical issues in accordance with the relevant protocols in force at the University of Ottawa, and indeed was selected in part to avoid such issues. I primarily based my evaluation with respect to ethics on the numerous peer-reviewed studies [e.g. Kramer et al., 2004; Holmer, 2008; Uthus & Aha, 2013], and at least two theses [Doell, 2000; Michels, 2012], that have used in a similar fashion as I do the logs of public IRC channels. Indeed, in one case, researchers used logs from the same Ubuntu-related source which I have analyzed myself, and prominently noted among the “advantages of using this corpus for research” that “all of the original chat messages are in the public domain” [Uthus & Aha, 2013, p. 99].

It is worth noting the following facts: anyone can join a public IRC channel, any contribution to such a channel is immediately available to all other participants, and all users can – and generally do – record or (in IRC parlance) “log” the collective conversations that take place there. Indeed, simply by virtue of using an IRC client, anyone can record IRC conversations, and as a matter of fact,

78 A “bot” in this context is a computer program designed to mechanically interact on an IRC channel.

79 A list of the channels thus selected can be found in 8.2 – Appendix 2: List of Publicly Logged IRC Channels Analyzed.

all major IRC clients automatically record conversations by default<sup>80</sup>.

Thus, much like on a platform such as Twitter, all messages written are, in effect, *public*<sup>81</sup>. For this reason, my role as an administrator of two channels studied is irrelevant with respect to the ethical aspects of my endeavour; this role only entailed that I was more familiar with these channels than others. Any user of these channels could access the same logs I have analyzed.

Nevertheless, for the purpose of preserving the privacy of the people studied, this work avoided publishing raw logs in part or whole, as a rule<sup>82</sup>. Furthermore, I eschewed the use of personally identifying information aside from the publicly available pseudonyms of participants, except when such pseudonyms were explicitly – and unequivocally publicly – associated with the corresponding legal identities<sup>83</sup>. The latter decision was deemed appropriate and scientifically valuable as it served to demonstrate the correspondence between the location of some participants' pseudonyms within diagrams, and these same participants' formal and public role within FOSS and/or community-driven projects (namely Ubuntu, MediaWiki, and the Wikimedia Foundation).

80 As an additional testament to the public nature of such IRC channels, websites exist that automatically and publicly display, and often keep track of, public channels' topics and users. For instance, clicking on any network listed on the page <https://search.mibbit.com/networks> yields a complete list of all public channels on that network, making available not only the current topic and number of users, but also the full *topic history* along with a *usage graph* for up to a year. The website <https://netsplit.de> offers similar information and features.

81 A very small number of public IRC channels have policies against the public sharing of logs, as a matter of courtesy. Much care was taken to ensure that this was not the case for any of the channels or groups of channels selected for analysis.

82 Making only three exceptions, respectively for the excerpts of text automatically generated by “bots” found in 8. 10 – Appendix 10: Additional Diagrams – Wikimedia-related channels, for the IRC interface sample (showing a glimpse of one of my own channels) found in 8. 15 – Appendix 15: IRC Interface Sample (Quassel Client), and for the highly impersonal spam sample found in 8. 16 – Appendix 16: IRC Log Sample – Spam in #ubuntu-offtopic.

83 In a few cases, I also directly or indirectly disclosed the country or state of residence of some pseudonymous participants. This was done with their unambiguous permission.

### **3. 5 – General Description of Diagrammatic Representations with KH Coder**

#### **3. 5 (a) – “Multi-Dimensional Scaling”:**

Such diagrams, as configured for the present study, consist primarily of circles of various colours, henceforth called “bubbles”, following *KH Coder*’s terminology. Each bubble stands for one of the *stemmed words* or *word-stems*<sup>84</sup> that are most common in the text represented, while its size increases proportionately to the frequency at which the paired word-stem occurs. In the legend situated at the right of the diagrams, the pertinent “Frequency” section includes 4 bubbles incrementally sized so as to illustrate the applicable frequency range of a given diagram, with the smallest and largest bubbles corresponding respectively to the least and most abundant word-stems included in that diagram. The number next to each bubble in the legend specifies the precise word-stem frequency associated with that bubble size.

For the purpose of striking a balance between comprehensiveness and simplicity, while bearing in mind the size limitations of this printable document, the “Multi-Dimensional Scaling” diagrams shown herein were made to include the  $40 \pm 1$ <sup>85</sup> most frequently occurring word-stems in the data-sets represented – omitting all “stopwords”<sup>86</sup> in all cases.

The distance between bubbles is generally commensurate with the degree of formal similarity between the statistical usage patterns of different word-stems, as measured through the Jaccard

84 During the automatic pre-processing of each data-set, all instances of words that were not already in their root form were reduced to that form by *stemming*, as described in [8. 3 – Appendix 3: Data Pre-Processing](#). The expression “word-stem” will henceforth be used when referring to such a shortened word.

85 41 word-stems were included when 40 could not be – due to more than one word-stem sharing a frequency equal to the threshold that would have resulted in precisely that many word-stems being selected. 39 word-stems were included when both 40 and 41 word-stems could not be selected for the same reason.

86 The entire list of which can be found in [8. 5 – Appendix 5: List of Stopwords Used](#).

similarity coefficient<sup>87</sup>. Colours shared by several bubbles represent “communities of ‘words’”, calculated by finding clusters of word-stems that (more or less) share an overreaching pattern of use. A number of such “communities” between 2 and 12 may be chosen; for the purposes of this research, the parameter used was 10, so that every cluster of bubbles contains 4 word-stems on average. The relevant part of the legend, entitled “Clusters”, lists all the clusters featured in a given diagram along with the colour codes that identify them; it also differentiates clusters by attributing to each of them a unique label, which simply consists of a numerical value determined according to an ordered sequence of integers starting at 1. These numbered labels are only intended for further reference, and do not otherwise hold any intrinsic meaning with respect to the content of the clusters they refer to.

The “dimensions” indicated on the *x*- and *y*-axes (labelled only *Dimension 1* and *Dimension 2* respectively), along with the sequence of numbers that serves as graduation for each axis, are abstract mathematical constructs that expressly serve the sole function of facilitating data visualization. How those “dimensions” and numbers relate to what they represent is essentially arbitrary: although they may sometimes serendipitously do so, they ordinarily *do not imply* that a diagram’s components are meaningfully arranged in any additional way apart from the features previously described pertaining to word-stem frequency and patterns of word-stem use. These mathematical constructs are strictly used to determine the *relative location* of a diagram’s elements (in this case, pairs of bubbles and word-stems), according to a manifold of statistical relationships that can be visually organized in a simplified model.

87 See the Wikipedia page: [http://en.wikipedia.org/wiki/Jaccard\\_index](http://en.wikipedia.org/wiki/Jaccard_index). A sample of a “Similarity Matrix” displaying a portion of the Jaccard coefficients for one data-set can be found in 8. 14 – Appendix 14: Sample of Similarity Matrix (Jaccard Similarity Coefficient).

A “Multi-Dimensional Scaling” diagram produced by *KH Coder* may display 1, 2 or 3 “dimensions”. However, in the present work, all such diagrams were intentionally generated with 2 “dimensions”, in view of that particular configuration resulting in what appeared to this researcher as more legible and uncluttered graphical representations of the data-sets analyzed – even ignoring the limitations inherent to the medium of paper.

Finally, a noteworthy advantage of a state-of-the-art “Multi-Dimensional Scaling” tool such as the one in *KH Coder* over other text visualization tools, for instance the popular “word clouds”, is that “Multi-Dimensional Scaling” diagrams that have been configured to include “bubble plots” (or any other equivalent feature) will display *both* word frequency and word association *at once*, whereas “word clouds” display *only* word frequency (as explained by Prof. Koichi Higuchi<sup>88</sup>).

### 3. 5 (b) – “Co-Occurrence Networks of Words”:

In this type of diagram, like in the one described previously, the approximate diameter of a circle or “bubble” (a “point” or “node” of the graph) increases in proportion to the frequency of the associated word-stem, so that the largest bubbles in diagrams of both types correspond to some of the most common word-stems in the data-sets they represent. In either type of diagram, the lowermost of the three parts of the legend at the right, entitled “Frequency”, illustrates the range of word-stem frequencies in the data-set and the precise frequency corresponding to bubbles of different sizes.

Unlike in the case of a “Multi-Dimensional Scaling” diagram which includes or excludes word-stems according to their frequency alone, the word-stems that ultimately appear in a “Co-

88 Personal communication [2014].

Occurrence Networks of Words” are only those *among* the most frequently occurring that *also* are part of the most common binary *word-stem-associations*. Therefore, “Co-Occurrence Network”-type visualizations of text data can be said to display the most significant binary conceptual associations and their mutual interconnections. Keeping in mind the same considerations with respect to the number of word-stems selected as for the other type of diagram (i.e. “striking a balance between comprehensiveness and simplicity”), “Co-Occurrence Networks of Words” created for the present research include salient *word-stem-associations* found among the  $40 \pm 1$ <sup>89</sup> most frequent word-stems, again with “stopwords” excluded, except when otherwise stated<sup>90</sup>.

While a “Multi-Dimensional Scaling” diagram does *implicitly* depict the statistical associations between word-stems – as each bubble’s position is determined relative to that of others according to the compared usage patterns of the associated word-stems – a “Co-Occurrence Network of Words” represents such statistical relationships *explicitly*. Thus, bubbles (“nodes” of the graphs) connected to one another by lines (“edges” of the graphs) represent word-stems occurring together especially frequently, either verbatim as a pair, or separated by a small number of other words (which are typically considered “stopwords” and thus excluded from analysis). The opacity of the lines codes for more frequently associated word-stems, measured using the Cosine measure of similarity<sup>91</sup>, which appears as a number displayed on each line; the “Coefficient” section of the legend at the right indicates the numerical coefficient corresponding to different opacity levels. The Cosine method was chosen in this case – instead of the Jaccard method used for “Multi-Dimensional Scaling” diagrams –

89 See footnote #81 in the previous section.

90 This is only the case for the diagrams in 8.11 – *Appendix 11: Additional Diagrams – Including Additional Verbs*, which include some common verbs omitted from other diagrams.

91 See the Wikipedia page: [https://en.wikipedia.org/wiki/Cosine\\_similarity](https://en.wikipedia.org/wiki/Cosine_similarity).

partly to provide a complementary perspective on the same text data, and partly because it was found to be better at highlighting features common to different data-sets.

Contrary to the structure of “Multi-Dimensional Scaling” diagrams, neither the absolute nor the relative position of bubbles holds any meaning within a “Co-Occurrence Network of Words”. The relationships between word-stems illustrated in such diagrams are entirely represented by the lines and coefficients that appear in the space between the bubbles.

Grey and *a fortiori* pink bubbles code for greater *centrality* of word-stems, a mathematical notion that purports to highlight the most semantically important nodes without relying on the simple measure of word-stem frequency. It can be computed in many different ways, ranging from a simple count of the number of nodes attached (in the case of “degree” centrality, not used here) to a count of the “frequency with which a point falls between pairs of other points on the shortest or geodesic path connecting them” [Freeman, 1978, p. 221] (in the case of “betweenness” centrality, used here).<sup>92</sup> The numerical values corresponding to different colours are shown within the “Centrality” section of the legend, at the top right of each diagram.

The most mathematically *central* word-stems, like the most frequent ones, tend to be found towards the centre of the graph. They are often more general in nature and are typically more likely to be found in analogous locations across multiple diagrams representing similar data-sets.

### 3. 5 (c) – Advantages and limitations of the diagrams:

In Peircean terms, *diagrams* such as the visualizations shown hereafter are defined as

92 See the Wikipedia page: <http://en.wikipedia.org/wiki/Centrality> and Freeman, 1978.

primarily *iconic* representations of assertions concerning states of affairs (in real or fictive universes)

[MS [R] 491: 2]<sup>93</sup>, or more precisely:

A *diagram* [...] is predominantly an icon of relations and is aided to be so by conventions. Indices are also more or less used. It should be carried out upon a perfectly consistent system of representation, founded upon a simple and easily intelligible basic idea. [CP 4.418]

In turn, *graphs* such as “Co-Occurrence Networks of Words” are a particular kind of diagram composed of points interconnected by lines which may or may not circumscribe bounded areas<sup>94</sup>.

Being icons, diagrams have a special epistemological status, well recognized by Peirce, who asserted that “Icons are specially requisite for reasoning “ [CP 4.531]. Indeed, the veritable constellations of meaning represented by complex diagrams offer the significant advantage of lending themselves readily to systematic logical analysis – many varieties of which (such as the lion’s share of mathematics) are altogether impossible without diagrammatic reasoning.

The specific types of diagrams included in this work offer the additional advantage of synthesizing mind-boggling amounts of information in a form that can be understood at a glance:

- The *smallest* data-set, used for [Figures 5a](#), [5b](#), and [V](#), spans 9,691 lines and 93,266 words.
- The *largest* data-set, used for [Figures 4a](#), [4b](#), and [IV](#), spans 582,416 lines and 5,800,369 words.

In total, the data-sets analyzed add up to nearly 2 million lines and 20 million words. Visualizing these text corpora within a dozen primary diagrams evidently constitutes a complex form of representation in and of itself, which would have been inconceivable without the aid of modern

93 Unlike ordinary language, which involves primarily *symbolic* representations of the same.

94 In Peirce’s words: “A *graph* is a superficial diagram composed of the sheet upon which it is written or drawn, of spots or their equivalents, of lines of connection, and (if need be) of enclosures.” [CP 4.419] Peirce’s notable “existential graphs” [CP 4.421] – diagrammatic representations of the elementary rules of logic, intended to facilitate complex logical operations – are a species of *logical graphs* (or *logical diagrams* [CP 4.347]), that is, graphs “representing logical relations iconically, so as to be an aid to logical analysis.” [CP 4.420]

computers and software. More specifically, each of these diagrams combines a *frequency table* with a *similarity matrix*<sup>95</sup> and several other mathematical measurements, such as the centrality of words<sup>96</sup>.

Yet, despite the many upsides of the types of visualizations chosen, it ought to be noted that the twin constraints of the software employed and of the format of a traditional book entail that the (static) diagrams of IRC conversations presented in this document cannot fully capture the *interactive, living* nature of *chat-rooms*. As a consequence, my conclusions regarding these aspects of the medium were drawn from direct observation, rather than the quantitative analysis of archived “logs” which resulted in the diagrams shown previously.

## **4 – Analysis**

### **4.1 – Sign Classification and Some Implications Thereof**

One of the most well-known and celebrated parts of Peirce's semiotic is his uniquely exhaustive classification of signs, developed by means of a combination of *trichotomies* conceived in accordance with his triadic definition of the sign and his three “phaneroscopic” categories of *Firstness*, *Secondness*, and *Thirdness*<sup>97</sup>.

“The most fundamental [division of signs] is into *Icons*, *Indices*, and *Symbols*” [CP 2.275], reflecting the way the sign relates to its dynamical object, in accordance with the three logical categories. While the more elaborate 1903 classification defines every class of signs through three

95 For example, [Figures 2a](#) and [2b](#) make use of the frequency table found in [8.13 – Appendix 13: Sample Table of Word Frequencies](#), and a similarity matrix akin to the one in [8.14 – Appendix 14: Sample of Similarity Matrix \(Jaccard Similarity Coefficient\)](#) but larger. That matrix is itself only one step in a computation that involves generating a table of *every paragraph of the document* crossed with every word selected.

96 As described in [3.5 \(b\) – “Co-Occurrence Networks of Words”](#).

97 This classification was detailed in [2.1 \(a\) – Peircean semiotic](#).

elements which amount to the relation of the sign to its ground and interpretant in addition to its object, it is not infrequent for the most fundamental division to be used alone “to represent the entire sign” [Liszka, 1996, p. 46].

Following the most fundamental division, two main types of signs are co-present in a typical IRC conversation. The most frequent is by far the *symbol*, as in any natural language-based interaction. We may infer that the meaning of an ordinary conversation unfolding on IRC depends chiefly upon convention or habit. [See EP 2:9] A second type of sign has a more subtle but nonetheless important presence: what I suggest be referred to as *symbolic icons* (or *symbolic likenesses*), namely signs – typically stylized and/or simplified – that simultaneously and inextricably display manifest features of icons *and* symbols (corresponding to *Secondness* and *Thirdness* with respect to how the sign relates to its object). Ignoring the self-evident presence of *computer icons* on every participant’s monitor, the most noticeable symbolic icons that are part of most text-based conversations on digital platforms nowadays are *emoticons*. Like other icons, by definition, emoticons represent their object through imitation (a simple “smiley” *looks like* a smiling human face), but their regular and conventionalized collective use to represent the concepts of various emotions has given them a concurrently symbolic character (unlike, for instance, in the case of a portrait or photo of a person smiling).

Following the 1903 classification, we primarily find in an IRC conversation three classes of signs, two of which ordinarily make up the greatest part of human language use. Alone capable of bearing truth or falsehood, and of being a vehicle for inquiry, these two uniquely human categories are: firstly, the “ordinary proposition”, classified as a *dicentic—symbolic—(legisign)*, or simply *dicent*

*symbol* [CP 2.262], and secondly, the structured discourse understood in the context of its relationship to a broader system of knowledge, which becomes an *argumentative*—(*symbolic*—*legisign*), or simply *argument*. As for the third class of signs, it once again primarily consists of emoticons: being simple “maps” of facial (or sometimes other bodily) features corresponding to the normal appearance of expressed emotions, but also obeying – to some extent, of course – conventions arbitrarily established by human beings<sup>98</sup> and relating to *emotions* considered as *general ideas*, they may be considered simultaneously as (*rhematic*)—*iconic*—*legisigns* (or *iconic legisigns*) and as *rhematic*—*symbolic*—(*legisigns*) (or *rhematic symbols*).

Those who claim that text-based media are “poor vehicles for emotion”, and that as a result text conversations are *low in emotion*, would do well to study the semiotic features of IRC interactions. Not only are the primary varieties of signs found in these interactions the same as the ones deliberately produced for any other conversation, but signs of a great diversity are used to express non-verbal cues normally present in face-to-face interactions<sup>99</sup>. Howard Rheingold, one of the first researchers to study IRC conversations systematically [1993 (2000)], appears to have reached much the same general conclusion in that sense. More recently, and perhaps more convincingly, a comprehensive review of research on the topic of emotion in computer-mediated communication found that “emotional communication online and offline is surprisingly similar, and if differences are found they show more frequent and explicit emotion communication in [computer-mediated

98 Such simple conventions may also be learned by many animals. Some apes such as chimpanzees, at least, have been shown to be capable of using the simplest symbols, categorized as *rhematic symbols*. However, the author of these lines has yet to see any convincing evidence for an animal mind having produced a single *dicent symbol* or *argument*. Computer algorithms in the “artificial intelligence” paradigm have had more (but still limited) success in this regard.

99 See for example the list of emoticons provided in [8. 17 – Appendix 17: Creative Use of Emoticons or “Smileys”](#), as well as the few examples of their use in context in [8. 15 – Appendix 15: IRC Interface Sample \(Quassel Client\)](#).

communication] than [face-to-face]” [Derks et al., 2008].

By contrast with this similarity, one major difference between text-based communication<sup>100</sup> and face-to-face communication which may be inferred from Peirce’s sign classification and its theoretical underpinnings is that the former type of communication will necessarily be less spontaneous (or, conversely, be more deliberate) than the latter. This is merely a necessary correlate of the absence of *rhematic indexical sinsigns* (e.g. a spontaneous facial expression or onomatopoeia) and other such *indices* which *involuntarily* represent such phenomena as human emotions by virtue of their *causal connection* with them.

The implications of this difference in the types of signs involved have fuelled numerous academic studies and debates – although the issue is seldom if ever conceptualized as such, but instead typically as a hypothetical difference in “authenticity” or “truthfulness”. While many argue that computer-mediated communication inherently lends itself to deceit and trickery (one author asserts that it “has become a growing hotbed for deception due to its provision of ubiquity, anonymity, and open environment” [Pak, 2014]), others conversely find evidence “that lying rates in face-to-face and instant messaging interactions were approximately the same” [Hancock et al., 2004], and more generally, that all “synchronous and recordless forms of media”, *including face-to-face communication* [Hancock, 2009], are generally the least representative of people’s genuine intentions.

In the view of the present author and according to his personal experience, there seems to be no good reason to consider it to be the rule rather than the exception for text-based communication (or, for that matter, computer-mediated communication) to be less honest than face-to-face

<sup>100</sup> Encompassing everything from an exchange of traditional letters to various forms of computer-mediated communication including, of course, IRC.

communication. It is worth remembering that con artists are often successful even when they meet their victims in person, while they are often unsuccessful on-line.

#### **4. 2 – #libertyinmind and #politics (KH Coder, Part 1)**<sup>101</sup>

##### 4. 2 (a) – Interpretation of *Figure 1a* and *Figure 1b*:

The simplest level of statistical text analysis being presumably a frequency comparison, I will begin each sub-section of the present chapter and the subsequent two by a simple list of the 15 most frequent word-stems that can be found in the preceding two diagrams (although only the corresponding “Multi-Dimensional Scaling” diagram, in this case Figure 1a, will necessarily include all of them<sup>102</sup>). In the present case, in decreasing order of frequency, these word-stems are:

“people”, “Aerides”, “lol”, “xd”, “try”, “law”, “mean”, “idea”, “read”, “Jigsy”, “mio-chan”, “point”, “argument”, “issue”, “interesting”

The single most frequent element, “people”, which also happens to be the most central one in Figure 1b, will be further discussed in a later section<sup>103</sup>, and will therefore be ignored for the time being. The second most frequent element, “Aerides”, is my own nickname, which it is all but surprising to see being prominently mentioned within the channel given how I acted as its creator and primary administrator. The third and fourth most frequent elements, “lol” and “xd” respectively, are signs that convey certain emotions: the first represents laughter, the second represents an ambiguous

101 The associated diagrams may be found in section 6. 1 – Diagrammatic Representations of #libertyinmind and #politics.

102 A “Multi-Dimensional Scaling” diagram, as opposed to a “Co-Occurrence Network of Words”, includes word-stems based on frequency alone, while the latter may exclude some frequently occurring word-stems that aren’t systematically paired with other commonly used ones.

103 In section 4. 2 (c) – Combined analysis of all diagrams representing #libertyinmind and #politics.

facial expression which can mean anything from “facetious affinity” to “amused disgust”. This suggests that the channel’s participants tended to express their own feelings frequently, and moreover that the feelings expressed were, more often than not, of a rather playful or joyous nature.

This seems especially plausible if we take into account the fact that most emoticons, such as “:-)” – often simplified to “:)” – “:- (“ – often simplified to “:(” – or “;-)” – often simplified to “;)” – were unfortunately culled from the text data by *KH Coder* due to being made up exclusively of punctuation signs. Extensive observation of the relevant channel over time has indicated to this student researcher that *at least some* among such signs would be included in the most frequent vectors of meaning used there, if only because “xd” is prominently featured in the diagrams, but did not in fact appear to be one of the emoticons that occurred most frequently in the conversations studied<sup>104</sup>.

Among the rest of the most frequent elements, “Jigsy” and “mio-chan” correspond to other frequent contributors to the channel’s conversations, while all other word-stems in the list (“try”, “law”, “mean”, “idea”, “read”, “point”, “argument”, “issue”, “interesting”) appear to be representative of what participants to #libertyinmind typically intended to do in that channel – the meaning of which I hope will be made clear in the remainder of this section. Two of the frequent word-stems just listed, “read” and “point”, are also respectively the 2<sup>nd</sup> and 3<sup>rd</sup> most central in [Figure 1b](#), which essentially indicates that they systematically occur across a great variety of contexts in this text corpus, and thus can be reasonably assumed to consist in a “common thread” linking together many of the channel’s conversations, by virtue of being crucial pieces of sign-equipment for the participants’ purposes.

104 Otherwise, the emoticon “xp” occurs as the 29<sup>th</sup> most common sign in the #libertyinmind data-set and is found in both [Figure 1a](#) and [Figure 1b](#), while “o\_o” is the 121<sup>th</sup> most frequently used sign in this data-set.

As will henceforth be done when perusing every subsequent “Multi-Dimensional Scaling” diagram, let us now dissect [Figure 1a](#) with an examination of the clusters that are most plainly meaningful with reference to my personal knowledge – extensive in this special case and that of #politics – of the conversations that took place in the relevant channel(s). This analysis, and every subsequent analogous part, will be supplemented whenever deemed appropriate by important features of the corresponding “Co-Occurrence Network of Words”, in this case [Figure 1b](#).

*Cluster 01*, located roughly in the centre relative to other clusters (accordingly, we might call it the *central cluster*, a useful expression for the reason that an analogous cluster is present in most other “Multi-Dimensional Scaling” diagrams shown herein), is the largest and perhaps the most significant among them. The most frequent word-stem, “people”, which may be viewed as the very centre of the diagram, is very closely associated with “idea”, only slightly less with “mean”, “argument”, “issue”, and “believe”, and slightly less still with “law” and “agree”. This seems to echo the channel’s general nature as an informal and congenial one dedicated to the free exchange of *meaningful ideas* and *arguments* relating to the *beliefs* about the *law* of its users, who largely *agree* with respect to these *issues*. Several word-associations in [Figure 1b](#) stand out which ostensibly support this interpretation, including three moderately strong ones: “people—believe” ( $C = 0,02$ ), “law—issue” ( $C = 0,02$ ), and “law—agree” ( $C = 0,02$ ).

*Cluster 04*, just below, seems to be the one closest in meaning to *cluster 01*, as it contains the nickname of the channel’s administrator “Aerides” along with “write” and “point”. Understanding that the users of #libertyinmind generally use the word “point” as a synonym for “argument”, it becomes clear – to the author of these lines, at least – that clusters *01* and *04* taken together are quite

representative of a channel having as its ultimate practical purpose the *writing* of polemical content about controversial legal topics (under the leadership of *Aerides*). That “agree—point” ( $C = 0,03$ ) is among the strong associations in the other diagram would seem to corroborate the existence of a conceptual link between the two aforementioned clusters.

*Cluster 05*, directly at the left of *cluster 01*, is perhaps the second most relevant among the clusters to understand the meaning of what was being discussed in the channel. Indeed, the four word-stems that it contains – “internet”, “speech”, “right”, and “drug” – all refer to precisely the kind of legal issues that the channel was created for people to discuss: the *rights* to freedom of *speech* and freedom of thought (the latter including “cognitive liberty”, which implies a limited right to use recreational *drugs*), especially as they relate to the *Internet* and the societal transformations it entails. The precise relationship between this cluster and *cluster 01* is substantiated by one of the strong word-associations found in [Figure 1b](#), “argument—speech” ( $C = 0,03$ ), while the moderately strong association “internet—right” ( $C = 0,02$ ) confirms that it is mostly people’s rights in cyberspace (and as they relate to cyberspace) to which these signs refer in this context, and “drug—law” ( $C = 0,02$ ) confirms that drugs are indeed discussed primarily with respect to the law.

*Cluster 08* groups only the three word-stems “interesting”, “reason”, and “child”. Of course, this is most likely in part because *interesting reasons* are given for various things on the channel – generally for the purpose of convincing others or being convinced, as illustrated by the moderately strong word-association “reason—agree” ( $C = 0,02$ ) in the second diagram. The inclusion of “child”, however, may well be coherently interpreted only by a person familiar with the channel at the time: indeed, during the period under consideration, one of the most contentious topics having piqued the

interest of the channel's users and being frequently debated was so-called "virtual *child* pornography" (pornographic material involving fictional characters resembling children, but never involving actual children in its computer-assisted production) and its implications for free speech on-line.

*Cluster 07*, while less semantically homogeneous, nonetheless brings together "try", "read", and "book", the combined meaning of which may be construed as relating to means of acquiring the knowledge needed to develop (and write) sound arguments. This can be illuminated by some of the word-associations in [Figure 1b](#): indeed, while the strongest association is by far "read—book" ( $C = 0,07$ ), we also find that "read—write" ( $C = 0,03$ ) is a strong association, hinting to the fact that the two activities are closely linked in the channel, while "try—write" ( $C = 0,02$ ) is a moderately strong association among those representative of the trial-and-error process unavoidable in the course of developing one's reasoning. We also find in *cluster 07* a second instance of "right", which necessarily corresponds to a very different pattern of use than the other instance. It is here positioned close to "xp", and to "xd" in *cluster 03* – both of which tend to be used somewhat separately from other signs rather than as normal parts of sentences – which suggests that the word is mainly used as a simple sign of assent in this context.

*Cluster 09* and *cluster 10* are the most opaque to my comprehension, but taken together, one finds that [Figure 1b](#) again sheds light on the context in which many elements are used: the pair of moderately strong associations "consider—fact" ( $C = 0,02$ ) and "consider—state" ( $C = 0,02$ ), thus, echo how the members of #libertyinmind frequently discuss *facts* pertaining to the role of the *state* in society (as well as relevant events that might have *happened*, to try to make sense of the word "happen" being included in *cluster 09*). In a similar fashion, "fact—life" ( $C = 0,02$ ) might be

understood as the two word-stems being jointly used in the context of debates such as those surrounding euthanasia, assisted suicide or abortion.

In *cluster 03*, “breadbasket”, an important member of the channel who participated extensively to discussions and happens to be a passionate video game enthusiast, is indeed found next to “game” – the very presence of which reflects the fact that most of the channel’s participants enjoy playing video games (among other kinds of games), and regularly discuss this hobby together. However, such conversations tend to be separate from others that occur in the channel, and less central. The only other sign included in this cluster is “xd”, which is however positioned somewhat apart, and therefore used less similarly, from the two previously mentioned.

Finally, in *cluster 11*, we find only the pseudonyms of two important participants at the time, namely “Jigsy” and “Miori”. It is worth noting that the creative monikers chosen by participants (“breadbasket”, “himarichan”, “Miori”, “Aerides”, “mio-chan”, “Jigsy”)<sup>105</sup> are always located towards the periphery of both diagrams, which reflects the fact that these symbols are used in a more unique way than ones which are in a more central location relative to others. Indeed, we also find “xd” and “lol” at the periphery of [Figure 1a](#), rather far from most other signs, which hints at the likely fact that these unconventional semantic units are used in a way somewhat distinct from ordinary words (although, by contrast, the emoticon<sup>106</sup> “xp” found in the lower right corner seems to be used more similarly to ordinary words).

The channel #libertyinmind being that of an informal activist organization in its early days, there is nothing surprising in the presence of “argument”, “believe”, “idea”, and “law” among the

<sup>105</sup> One might think of them as “stage names”, in accordance with Goffman’s conceptual framework.

<sup>106</sup> An “emoticon” is composed of text characters, while an “emoji” is an actual picture that serves the same purpose.

elements of the central cluster, nor in the presence of the strong word-associations “read—write” ( $C = 0,03$ ), “agree—point” ( $C = 0,03$ ), and “argument—speech” ( $C = 0,03$ ) in [Figure 1b](#): the main activities which the users participate in for the purposes of the channel are by far *reading* and *writing* to develop *ideas* and *arguments* about *points of agreement* in participants’ *beliefs* about *laws* relating to freedom of *speech*. “In semiotic terms, such words or symbol [replicas] can be construed as interpretants of activism or the aim to organize it through that medium”<sup>107</sup> – making use of Wittgensteinian terminology, we might speak of a *family of interpretants*. This family can be said to include all signs in the diagrams, with the possible exception of pseudonyms, “lol” and emoticons.

#### 4. 2 (b) – Interpretation of *Figure 2a* and *Figure 2b*:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“people”, “lol”, “Aerides”, “himarichan”, “state”, “mean”, “view”, “try”, “agree”, “child”, “drug”, “believe”, “government”, “political”, “point”

The single most frequent word-stem, “people”, which is again the most central in [Figure 2b](#), will be discussed in detail in the next section<sup>108</sup>, and will therefore be disregarded for now. It surely is unnecessary to repeatedly explain the meaning of “lol”<sup>109</sup>, but here again, the expression being found so far up the list of most frequently used symbols would seem to suggest that the channel’s users frequently express feelings of amusement in the context of their conversations there. “Aerides” and

107 Fernando Andacht, personal communication [2015].

108 In section 4. 2 (c) – Combined analysis of all diagrams representing #libertyinmind and #politics.

109 Arguably, it may well have been unnecessary in the first place. The fact that few people – in developed, Western societies at least – are unaware of the expression is an excellent example of the changes that occur within our language-games in this day and age *due to the influence of new communication technologies*.

“himarichan” simply correspond to the two administrators of the #politics channel during the period under consideration. Most of the remaining elements included in the list (“state”, “mean”, “view”, “try”, “agree”, “believe”, “government”, “political”, “point”) intuitively represent concepts to which one might normally have recourse in the majority of discussions of a political nature. Two exceptions, which seem oddly specific by comparison, are the word-stems “child” and “drug”.

The presence of the word-stem “drug” among the most frequent would appear to be one of a few rare instances, in the text corpora analyzed, where outside events had a clear and important impact on the most common signs used. Indeed, while #politics was regularly home to discussions of the so-called “war on drugs”, this became especially prominent during the period under scrutiny due to the legalization of cannabis, which was beginning to gain significant political momentum in many areas of the world, including several (such as Colorado, California, Canada, and New Zealand) where some of the channel’s participants lived. In fact, “himarichan”, the other administrator of the #politics channel at the time apart from myself, hailed from Colorado, one of the first two American states (along with Washington) to legalize the recreational use of cannabis in 2012. Furthermore, both he and myself were vocal in our support of that particular cause.

As for the word “child”, the fact that it is the 10<sup>th</sup> most common word-stem in the list above is slightly more puzzling at first glance. A plausible hypothesis emerges from a comparison with #libertyinmind, considering that several active participants in one channel were also active in the other, and that both channels were recorded during a similar period: as mentioned in the previous section, at the time, the topic of “virtual *child* pornography” was being regularly debated by several people participating in both channels (who often used them to discuss the same issues with different

people and from different angles). If we add to this the fact that #politics was also frequently home to discussions on multiple other topics, such as education, involving the place of children within society, it becomes all but surprising that this word-stem would be so common.

Our examination of the most meaningful clusters in [Figure 2a](#) may thus begin with *cluster 01*, which includes the word-stems “people”, “believe”, “idea”, “child”, and “drug”. The first three word-stems listed can be interpreted as the most general theme describing the conversations taking place in the channel: in other words, *people* join #politics to discuss their *beliefs* and *ideas*. The last two word-stems, for their part, may be understood as referring to controversial topics which were among the ones most often debated in the channel during the period analyzed, for the reasons explained above.

*Cluster 05* and *cluster 07* are arguably the most representative of the channel’s specific and distinctive purpose: they include “Aerides”, the pseudonym of the channel’s primary administrator<sup>110</sup>, and “pimentel28”, one of the most frequent contributors, but more significantly, they also include “political”, “view”, “tell”, “channel”, and “politics”. This seems to accurately describe a *channel* about *politics* where people *tell* others about their *political views*. Indeed, in [Figure 2b](#), we find that “political” is the second most central word, strengthening the notion that this is *actually* what most conversations on #politics have to do with. In addition, “political—view” ( $C = 0,14$ ) is by very far the strongest word-association, while we also find the moderately strong association “political—channel” ( $C = 0,03$ ), the latter appearing to corroborate the relevance of grouping clusters *05* and *07* together.

If we have now firmly established the general subject matter of #politics, as well as some momentary topics being prominently debated at the time, *cluster 03* encompasses – along with

<sup>110</sup> Also notable is the fact that it is located somewhat more towards the middle than the other three pseudonyms in [Figure 2a](#), “himarichan”, “lpf”, and “pimentel28”, indicating that it occurs in more diverse contexts.

“himarichan”, the channel’s second administrator – four word-stems which can be said to represent more concretely *what participants do* and *why they do it*: “interesting”, “point”, “agree”, and “read”. That is to say, those who take part in discussions in the #politics channel argue by making *points*, hopefully *interesting* ones that would get others to *agree*, and they *read* what was written (or shared via hyperlink) by others in case they would themselves find it *interesting* and/or *agree*.

*Cluster 08*, next to *cluster 05*, comprises the three word-stems “speech”, “social”, and “issue”. The last two are probably best considered as a pair, like they occur in the form of a moderately strong link in [Figure 2b](#): “social—issue” ( $C = 0,03$ ). The expression “social issue” may be better understood in reference to some theories in political science – upon which is based the on-line *Political Compass* tool<sup>111</sup>, frequently mentioned in #politics – which divide the spectrum of political views according to two axes, respectively *social issues* and *economic issues*. Topics related to the former axis were more often discussed in the channel than ones related to the latter axis, likely first and foremost because many users were also active in #libertyinmind where they engaged in activism pertaining to social issues such as freedom of *speech* – which is manifestly why this same cluster also includes “speech”. That “social issues” may be adequately interpreted in light of the intricacies of “political views”, and thus that there must be an important conceptual link between *cluster 08* and *cluster 05*, is further substantiated by the existence in the other diagram of the very strong word-association “political—issue” ( $C = 0,04$ ) and the moderately strong one “social—view” ( $C = 0,03$ ).

*Cluster 04*, *cluster 06*, and *cluster 11* all ostensibly consist of concepts that are likely to be discussed as part of various political debates. *Cluster 04* thus includes the word-stems “state”,

111 See: <https://www.politicalcompass.org/>.

“system”, and “live”, words which seem likely to be used in the context of conversations about the political *system* (a word which is also the third most central in [Figure 2b](#)) of the *state* in which a person *lives*.

In *cluster 06*, the word-stems “government”, “life”, “world”, and “money” barely require any explanation of their relevance to politics: the role of *government* practically defines the word, the role of *money* is of paramount importance to political matters such as power dynamics in general or taxes in particular, the concept of *life* is crucial to various political debates such as those relating to euthanasia, assisted suicide or abortion (all of which were also often discussed in #libertyinmind, as was previously mentioned), and the presence of *world* in the cluster echoes the international character of the channel’s users and conversations. On that note, it is also noteworthy that a very strong association exists between the two words “government—world” ( $C = 0,04$ ), from which we might infer that the place of government within society tended to be discussed from an international perspective – and that the idea of *world government* is likely to have been debated in the channel.

In *cluster 11*, we come across three word-stems: “right”, “problem”, and “country”. To all appearances, the first probably corresponds mainly to the notion of human *rights*. The second evidently refers to the multiple perceived *problems* within society that any proposed political solution necessarily seeks to address. The third simply means that conversations in #politics frequently invoke actual nation-states, either because of current events happening in a given *country* or because a participant lives there (indeed, it is no coincidence that “country” is located close to “live” and “state” in *cluster 04*).

*Cluster 10*, which includes “consider”, “law”, and “happen”, might be interpreted according

to the knowledge that *considering* what might *happen* if we changed the *law* is a critical part of political reasoning. More generally, it is to be expected that political discussions will involve considering the law, what has happened, and what could happen. The fact that “happen” is positioned somewhat away from other elements in [Figure 2a](#) indicates that it is typically used in a relatively distinctive fashion – what this fashion is, however, is not immediately clear.

Other signs positioned in a similar fashion as “happen” are “xd”, which is the sole bubble in *cluster 12* and thus unmistakably amounts to a rather unique pattern of use, and “lol”, in *cluster 02*. It is worth remembering that the same two signs were found in analogous locations in [Figure 1a](#), the significance of which will be pondered further in the next section.

In the same way as in #libertyinmind, we find a *family resemblance* in the elements of the diagrams representing #politics. These elements may be said to consist in a family of interpretants of political discussion and debate.

#### 4. 2 (c) – Combined analysis of all diagrams representing #libertyinmind and #politics:

The overlap between the *#libertyinmind* (2013-10-29 to 2014-08-04) and *#politics* (2012-04-29 to 2014-08-04) datasets was expected to be extensive as the channels shared the same administrator and part of their user-base, in addition to having been recorded simultaneously after October 29<sup>th</sup>, 2013 (although #politics started being recorded before that date).

Already, we find substantial similarities between the most frequently occurring elements in each channel: “people”, “Aerides”, “lol”, “try”, “mean”, and “point” are found among the 15 most

common word-stems in both data-sets. Furthermore, a quick look at [Figure I](#) and [Figure II](#) – “Co-Occurrence Networks of Words” similar to [Figure 1b](#) and [Figure 2b](#), but with some additional verbs taken into consideration – reveals that “think” and “should” are respectively the 2<sup>nd</sup> and 3<sup>rd</sup> most frequent words in both cases. A three-way association between “people”, “think”, and “should” seemingly constitutes a distinguishing feature of the two channels, which might echo the fact that both channels are primarily dedicated to what *people think*, and that political activism and political debates generally pertain to what *should* be the case.

The reason for “people” being – by a fair margin – the most frequent and most central word-stem is far from self-evident. One might assume that it is merely by virtue of being one of the most common English words in general, which is surely the case<sup>112</sup>, but this does not explain why it is only in the *#libertyinmind* (2013-10-29 to 2014-08-04) and *#politics* (2012-04-29 to 2014-08-04) data-sets, and in no other data-set among the ones analyzed for the purposes of this study<sup>113</sup>, that it is *more frequent than all other word-stems*. It seems plausible that the word is more frequent and central in more casual channels where users tend to engage in socializing, a hypothesis which will be examined again later<sup>114</sup>.

Another major similarity between both data-sets is illustrated by “people”, “believe”, and “idea” being part of the central cluster in both “Multi-Dimensional Scaling” diagrams, which is representative of how both channels are home to conversations that predominantly feature *people’s beliefs* and *ideas*. Indeed, a fact corroborating this interpretation is that the only word-association

112 The *Oxford English Corpus* ranks it 61<sup>st</sup> among the 100 most common words [see: [https://en.wikipedia.org/wiki/Most\\_common\\_words\\_in\\_English](https://en.wikipedia.org/wiki/Most_common_words_in_English)].

113 Encompassing the data-sets that were ultimately included in this thesis along with many more, of multiple varieties, that this student researcher experimented with over the course of this research.

114 In section [4.3 \(d\) – Interpretation of Figure 6a and Figure 6b](#), pertaining to diagrams representing #ubuntu-offtopic.

which has a coefficient above  $C = 0,01$  in both [Figure 1b](#) and [Figure 2b](#) is “people—believe”.

Nevertheless, a remarkable point that both [Figure 1a](#) and [Figure 2a](#) have in common is how the signs “lol” and “xd” are located in analogous locations at the periphery. This plausibly indicates that the same vocabulary of creative, non-grammatically standard neologisms is regularly used within both channels *in very much the same way*. Moreover, precisely how these signs are used is systematically different from the way ordinary words are used, which could simply mean that they are generally used alone as “messages” in and of themselves. In a similar but less pronounced fashion, the fact that nearly all pseudonyms found in these same two diagrams – with the sole exception of “Aerides” in [Figure 2a](#) – are positioned towards the periphery rather than the centre reflects the distinctiveness of the way each of them is generally used.

In addition, despite differences in their relative frequency and in how they are clustered, more word-stems appear in both “Multi-Dimensional Scaling” diagrams than word-stems appearing in only one of them. Signs which appear in both are listed below (in alphabetical order):

“Aerides”, “agree”, “believe”, “channel”, “child”, “consider”, “drug”, “happen”, “himarichan”, “idea”, “interesting”, “issue”, “law”, “life”, “lol”, “mean”, “part”, “people”, “point”, “read”, “right”, “speech”, “state”, “try”, “xd”

By contrast, the following are found exclusively within the diagrams that represent each channel:

- *#libertyinmind*: “argument”, “book”, “breadbasket”, “fact”, “game”, “internet”, “Jigsy”, “mio-chan”, “Miori”, “reason”, “sense”, “watch”, “write”, “xp”
- *#politics*: “country”, “government”, “live”, “lpf”, “money”, “pimentel28”, “political”, “politics”, “problem”, “social”, “system”, “tell”, “view”, “war”, “world”

The two lists above illustrate the specificity of each channel in contrast to the other – and,

accordingly, the *shared purpose* which defines each of the two channels' users. The fact that the signs included in the list corresponding to #politics appear to be more specifically representative of the channel's explicit subject-matter reflects how conversations in #libertyinmind are more casual than those in #politics, consistent with my personal experience participating in the two channels.

#### **4.3 – Ubuntu-related channels (KH Coder, Part 2)**<sup>115</sup>

##### 4.3 (a) – Interpretation of Figure 3a and Figure 3b:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“install”, “try”, “ubuntu”, “file”, “package”, “run”, “thanks”, “problem”, “sudo”, “tell”, “help”,  
“linux”, “server”, “error”, “mount”

The most frequent, and most central, word-stem is thus “install”, a technical term with a specific use in the context of computing, like most other elements listed above (“ubuntu”, “file”, “package”, “run”, “sudo”, “linux”, “server”, “error”, “mount”) – the meaning of which will be gradually explained throughout this section. If we ignore such words, however, the remainder of the list already starts to paint a more intuitively understandable picture of the purpose of the channels in this data-set: “try”, “thanks”, “problem”, “tell”, and “help” appear like they could well describe how people typically join these channels to *tell* others about their *problems* in the hope that someone will *try* to *help*, success in doing so being rewarded with a *thanks*.

Indeed, *cluster 02* includes the word-stems “thanks” and “help”, the latter being located next

<sup>115</sup> The associated diagrams may be found in section 6.2 – Diagrammatic Representations of Ubuntu-related channels.

to “please” (in *cluster 11*), which we also find in [Figure 3b](#) within the very strong word-association “please—help” ( $C = 0,06$ ). We find a second instance – possibly because one is categorized as a noun and the other as a verb – of the word-stem “help” next to “people”, both in *cluster 06*, which also includes “lol”, albeit it is located rather apart from the cluster’s (and the diagram’s) other elements.

What is it, though, that people request help with? We may now turn our attention towards other meaningful clusters in [Figure 3a](#), starting with the central cluster, *cluster 01*. This cluster, which includes the six most frequent word-stems in this data-set, may be divided into *two triads of contiguous elements*. The first, perhaps the most meaningful with respect to the question above, comprises “install”, “ubuntu”, and “package”, and is therefore analogous to the combined presence of “install—ubuntu” ( $C = 0,06$ ) and “install—package” ( $C = 0,06$ ) among the very strong word-associations in the other diagram. The second includes “try”, “run”, and “file” – the first two words being also fairly central – which we may interpret as referring to how those who seek assistance on such channels are likely to be told to *try* various things, a critical part of which will probably be to *run* (meaning “execute” in computing jargon) some programs by making use of the relevant *files*.

By now, the reader might have guessed that *packages* are the way the Ubuntu operating system, as with most modern GNU/Linux distributions, typically *installs* additional software, and that the majority of users who join the more active Ubuntu-related channels apparently want to know how to *install packages*, if not how to *install Ubuntu* itself.

*Cluster 03*, encompassing the word-stems “problem”, “linux”, “window”, “server”, and “user”, can be understood as describing the fact that the *problems* being addressed within these

channels are experienced by *users* of the “*Linux*” operating system<sup>116</sup> (Ubuntu being a variety thereof, as highlighted by the proximity of “linux” and “ubuntu” in this diagram) and/or the *Windows* operating system (the word-stem “window” in this cluster is most likely a result of the stemmer shortening the word “Windows”), either of which may be used as a *server*. The hypothesized conceptual link between “Linux” and “Windows” is corroborated by the fact that “linux—window” ( $C = 0,07$ ) is among the strongest word-associations in [Figure 3b](#), but as we might expect given the nature of the text corpus, only “linux” out of these two word-stems is also specially central (although not extremely so). However, in light of the very strong word-association “add—user” ( $C = 0,06$ ), it is worth noting that the concept of “user” certainly refers here both to its usual, everyday meaning and to its technical meaning with respect to operating systems (it is only in this latter sense that one might, strictly speaking, *add a user*).

*Cluster 09*, corresponding to the word-stems “tell”, “mean”, “change”, and “source”, appears rather semantically heterogeneous at first glance. If nothing else, it seems reasonable to think that people experiencing problems with their operating systems would be asking others to *tell* them – to *please* tell them, reflecting the relatively weak but significant association “please—tell” ( $C = 0,03$ ) – what the issues encountered *mean*. Solutions consequently offered surely involve some *changes* to the software on one’s computer, which for an Ubuntu user might entail an operation such as *adding package sources*, as illustrated by the word-associations “add—source” ( $C = 0,07$ ) and “package—source” ( $C = 0,05$ ), respectively one of the strongest and a strong one found in the “Co-Occurrence Network of Words”, both exclusively so.

116 The term “Linux” is commonly used to refer to the complete GNU/Linux operating system, such as when contrasting it with Windows, although it technically refers only to the *Linux kernel*.

In addition to these words being relevant to the context of technical assistance, they are also useful within the context of *software development* (involving modifying the *source code* of computer programs<sup>117</sup>), which we might reasonably expect to be another common purpose of the users of the channels under analysis. This purpose will be further described in later sections.

*Cluster 04* includes the three elements “sudo”, “command”, and “type”: *sudo*, used to obtain temporary administrative privileges, is very commonly *typed* at the beginning of *commands* (in the text-based “command-line”, or “console”, user interface) on Ubuntu and similar operating systems. The importance of the *sudo* command is further highlighted by the high centrality (respectively third and second in rank) of the two word-stems as they appear in [Figure 3b](#). *Cluster 05* groups together three elements as well, “error”, “read”, and “mount”, which can be parsed as relating to two types of technical issues which may be encountered on a computer, namely *mount errors*<sup>118</sup> and *read errors*<sup>119</sup>.

*Cluster 08* is primarily composed of three highly technical terms, “card”, “driver”, and “kernel”: the first describes a broad type of computer hardware component (more specific categories of which are, for instance, graphics *cards* and network *cards*)<sup>120</sup>, the second is a kind of software that enables the operating system to use specific devices (thus, a particular *driver* can make it possible to use a particular printer)<sup>121</sup>, and the third is a component of an operating system that serves as its core

117 The defining feature of “open source” projects such as Ubuntu or MediaWiki is that their source code is provided for free for anyone to see, as opposed to proprietary software which typically implements various means to obfuscate the source code for the purpose of protecting their intellectual property.

118 This kind of error concerns how an operating system symbolically integrates into its own file system, and makes available to its users, the files and directories on a storage device such as a hard drive, DVD, or Flash drive. See the Wikipedia page: [https://en.wikipedia.org/wiki/Mount\\_\(computing\)](https://en.wikipedia.org/wiki/Mount_(computing)).

119 This kind of error occurs during the process by which a device such as a hard drive or CD/DVD reader physically “scans” a source to retrieve data. See the Wikipedia page: [https://en.wikipedia.org/wiki/Reading\\_\(computer\)](https://en.wikipedia.org/wiki/Reading_(computer)).

120 See the Wikipedia page: [https://en.wikipedia.org/wiki/Expansion\\_card](https://en.wikipedia.org/wiki/Expansion_card).

121 See the Wikipedia page: [https://en.wikipedia.org/wiki/Device\\_driver](https://en.wikipedia.org/wiki/Device_driver).

and oversees all its processes (an example of which is the Linux *kernel*)<sup>122</sup>. Two strong word-associations in the other diagram, “install—driver” ( $C = 0,05$ ) and “kernel—source” ( $C = 0,05$ ), provide examples of the contexts in which these words are commonly used in this text corpus. In addition to the three word-stems previously mentioned, *cluster 08* also contains the only pseudonym included in these diagrams, “Kyrall”<sup>123</sup>, which is located in a corner, quite apart from most other elements in Figure 3a.

*Cluster 07*, finally, includes three word-stems that are worthy of mention: “system”, “version”, and “breezy”. While the last word’s inclusion might initially appear strange to the reader, everything should become clear after I explain that *Breezy Badger* was the name given to the most recent *version* of the Ubuntu operating *system* as of December 25<sup>th</sup>, 2004.

#### 4. 3 (b) – Interpretation of Figure 4a and Figure 4b:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“install”, “try”, “run”, “thanks”, “ubuntu”, “file”, “package”, “problem”, “bug”, “issue”, “add”, “test”, “mean”, “server”, “error”

In the same way as for the previous data-set, it is possible to split the list above into two parts according to whether or not each element has a specific technical meaning in this context. The most frequent word-stem is once more “install”, although it is now “only” the second most central. If we ignore other words which have a specific use in computing (“run”, “ubuntu”, “file”, “package”,

122 See the Wikipedia page: [https://en.wikipedia.org/wiki/Kernel\\_\(operating\\_system\)](https://en.wikipedia.org/wiki/Kernel_(operating_system)).

123 Corresponding to Chris Peterman, an ex-contributor to the Ubuntu project. [<https://launchpad.net/~kyral>]

“bug”, “test”, “server”, “error”), we are left with the following: “try”, “thanks”, “problem”, “issue”, “add”, and “mean”. While the first three were identical in the previous case, the remainder is unique to the most frequent words in this data-set. The word-stem “issue” can be simply understood as synonymous with “problem”, but interpreting the kind of contexts in which “add” and “mean” tend to be used here proves to be somewhat more challenging.

Let us now take a closer look at the clusters in the “Multi-Dimensional Scaling” diagram, Figure 4a. *Cluster 03* includes the three most frequent word-stems, “install”, “try”, and “run”, which are analogous in meaning to the same terms as described in the previous section (in addition, in a way not too dissimilar from what was observed in the previous data-set, “run” and “install” are respectively the most central and second most central words). However, the same cluster also comprises the word-stems “file”, “system”, “error”, and “server”, all of which ought to be interpreted according to their technical meaning: a person might encounter *errors* with their operating *system*, which may or may not be used to run a *Web server*<sup>124</sup>. The fact that Ubuntu is commonly used in that way is substantiated by the strong word-association “ubuntu—server” ( $C = 0,05$ ), while either the problem or potential solutions will probably relate in some way to the *files* on the person’s computer.

Directly below, we find *cluster 02*, which groups together “ubuntu”, “problem”, “Ubuntu”, and “14.04”. The word-stem “problem” refers to how users of these channels overwhelmingly join them for the purpose of requesting technical assistance – to find solutions to their *problems* with the *Ubuntu* operating system. The numbers “14.04” simply correspond to the most recent version of Ubuntu at the time, which is confirmed by the very strong associations that exist between “ubuntu—

124 See the Wikipedia page: [https://en.wikipedia.org/wiki/Web\\_server](https://en.wikipedia.org/wiki/Web_server).

14.04” ( $C = 0,06$ ) and “Ubuntu—14.04” ( $C = 0,06$ ). It is also notable that, in much the same way as in [Figure 3b](#), we find the pair “install—ubuntu” ( $C = 0,05$ ) among the strong word-associations in [Figure 4b](#), which echoes the fact that many of the people who use Ubuntu-related channels do so for the purpose of *installing* the *Ubuntu* operating system.

*Cluster 09*, including the word-stems “help”, “please”, “question”, and “tell”, reflects the process of asking for technical assistance which manifestly describes the majority of conversations on Ubuntu-related channels. In the “Co-Occurrence Network of Words”, we find the signs “please”, “help”, and “tell” detached from others and interconnected moderately strongly as “please—help” ( $C = 0,04$ ) and weakly as “please—tell” ( $C = 0,02$ ).

*Cluster 04* includes elements – “package”, “add”, “update”, “test”, “app”, and “fail” – which ostensibly relate to practical solutions that might be suggested to a user’s problems. Indeed, we might expect that problems with the Ubuntu operating system will often be solved by *updating* software *packages* (which correspond to *apps*) or *adding* new ones by installing them – as is illustrated by the strongest word-association in [Figure 4b](#), “install—package” ( $C = 0,07$ ). In the process, it is likely that extensive *testing* will be required – which is mirrored, in technical language, by the strong word-association “run—test” ( $C = 0,05$ ) – and that many such attempts will probably *fail* – which is similarly mirrored by the strong word-association “test—fail” ( $C = 0,05$ ).

*Cluster 08* is similar in meaning to *cluster 04*, that is, it concerns actions that may be done on someone’s operating system. This is best illustrated, as in the previous data-set, by “sudo” and “command”, which again refer to a very common way to begin commands on Ubuntu and similar operating system, when administrative privileges are temporarily needed. The fact that “sudo” is

located quite apart from other elements in [Figure 4a](#), reflects how it is distinctively used as part of commands rather than as part of ordinary sentences.

To give one example, if an Ubuntu user wishes to install the “firefox” package – which is, of course, the *Firefox* web browser – one might, although there are other ways to achieve the same result, type and run the following command (unsurprisingly, the pair of word-stems “run—command” ( $C = 0,05$ ) are strongly associated with each other):

```
sudo apt install firefox
```

Similar commands used to install software appear to be especially common in this data-set, as we discern a strong association between the two signs “sudo—install” ( $C = 0,05$ ).

The other elements included in *cluster 08* are “set”, “user”, and “remove”. The word “set” is a polysemous one<sup>125</sup>, but it is likely that its primary use in this data-set is as a verb, meaning roughly “put or bring into a specified state”<sup>126</sup>, which applies to a great variety of situations in the matter of computing, most evidently those related to a computer’s settings. As for “user”, it mostly refers here to its technical use in this context, such as when an operating system allows different *users* to have different settings, separate personal files, etc. The word “remove”, for its part, is also relevant to countless possible operations that may be done on a computer, but two examples that involve other words in the same cluster are to *remove* a *user*, and most importantly the *command* to uninstall a package, which would be, in the case of “firefox”:

```
sudo apt remove firefox
```

*Cluster 05* includes the word-stem “right”, “bug”, “fix”, and “build”. While the first word-

125 See, for instance, the dictionary entry: <https://www.merriam-webster.com/dictionary/set>.

126 This is definition #2 provided by Google: <https://www.google.com/search?q=define+set>.

stem “right” is probably merely used as a symbol of assent in most contexts in this data-set<sup>127</sup>, the other three all relate first and foremost to *software development*. One of the most crucial parts of improving computer software is indeed to *fix bugs*, while the goal of writing software within the Ubuntu ecosystem is to *build* packages from the source code to include them in the software repositories where the operating system’s users can get the programs they want.

*Cluster 06* can be interpreted both in light of the process of software development and in light of computer use generally. The word-stem “issue” is simply synonymous with “problem”, which is reflected in the fact that the two corresponding bubbles are next to each other in [Figure 4a](#). The word-stem “version” is germane to the creation of new *versions* in the course of improving software, and equally so to the fact that such new versions will ultimately be installed by end-users in the process of updating their software (it is no coincidence that “update” is found next to “version”). The word-stem “check”, for its part, is simply used here as a synonym for “verify”, which in the context of computing is part of making sure that software works as intended, both by humans while developing said software, and automatically by the software itself (such a verification mechanism is used, for instance, every time the operating system starts).

*Cluster 07* includes the two word-stems “window” and “boot”. The first might refer at once to the windows that make up ordinary graphical user interfaces (GUIs) on modern computers, and to the *Windows* operating system – as explained in the previous section, it is very likely that the stemmer mistakenly shortened the word because people frequently write it without an initial capital letter. The second evidently refers to the process of starting up a computer, and its inclusion in the same cluster

127 As echoed by its position at the periphery of the diagram, as previously explained in section [4.2 \(a\) – Interpretation of Figure 1a and Figure 1b](#).

as “window” probably reflects in part how many users of Ubuntu install it along with Windows on the same computer, something called “dual-booting”.

Finally, two clusters comprise one element each: “thanks” for *cluster 01* and “Saviq”<sup>128</sup> – the only pseudonym in these diagrams – for *cluster 11*. These two word-stems manifestly correspond to patterns of use that are particularly distinctive, such as “thanks” being regularly used alone as a message in and of itself.

#### 4.3 (c) – Interpretation of *Figure 5a* and *Figure 5b*:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“install”, “try”, “kubuntu”, “file”, “problem”, “run”, “system”, “package”, “BluesKaj”, “driver”,  
“KUBUNTU”, “thanks”, “KDE”, “ubuntu”, “window”

This data-set corresponds to a single channel also included in the previous data-set, although the time-frame selected is quite different (January 2013 as opposed to July and August of 2014<sup>129</sup>). Nevertheless, it would seem that Figures 5a and 5b are most similar to Figures 4a and 4b.

Again, the two most frequent word-stems are respectively “install” and “try”, with “install” being once more the second most central element in Figure 5b. The third most frequent word is “kubuntu”, which is to say the least unsurprising for channel called #kubuntu. The Kubuntu operating system<sup>130</sup> is a version of Ubuntu which uses the KDE desktop environment instead of the one used by

128 Michał Sawicz, member of many teams related to the (largely volunteer-driven) programming of the Ubuntu operating system. [<https://launchpad.net/~saviq>]

129 These time-frames were randomly selected, with the only restriction being that either one or two months were to be included in each case.

130 The author of these lines uses Kubuntu as his primary operating system. It is for that reason that the #kubuntu

default, GNOME – this is mirrored in the fact that “kubuntu” and “ubuntu” are very close to each other in *cluster 01*, and close to “KUBUNTU” in *cluster 06*; similarly, it is reflected in the strong word-association “Ubuntu—KUBUNTU” ( $C = 0,06$ ).

I will now offer a plausible detailed interpretation of each sufficiently meaningful cluster in [Figure 5a](#). *Cluster 07*, the central cluster, includes the word-stems “install”, “try”, “run”, and “desktop”, and is partly analogous to both *cluster 01* from [Figure 3a](#) and *cluster 03* from [Figure 4a](#). More specifically, the first three word-stems listed have a function highly similar to that of their counterparts appearing in the previous two “Multi-Dimensional Scaling” diagrams analyzed. In the diagram under consideration, however, *cluster 07* contains the additional word-stem “desktop”, presumably highlighting the fact that the Kubuntu operating system has as defining feature its *desktop* environment, which seems to be an important and recurrent subject of the discussions in the channel.

*Cluster 01*, which encompasses the word-stems “ubuntu”, “kubuntu”, and “package”, is located very close to *cluster 07*, and thus corresponds to similar patterns of use – the word-stem “install” even overlaps “kubuntu”. Indeed, what users generally talk about *installing* are the *Kubuntu* or *Ubuntu* operating systems, as well as *packages*, as was previously discussed in the two preceding sections about Ubuntu-related channels. The nature of the conceptual bond between *cluster 07* and *cluster 01* is further corroborated by the fact that “install—KUBUNTU” ( $C = 0,08$ ), “install—kubuntu” ( $C = 0,07$ ), and “install—package” ( $C = 0,07$ ) are strong to very strong word-associations in the other diagram.

*Cluster 06* brings together the signs “KUBUNTU”, “12.10”, “version”, and “mean”. For the channel was selected for analysis, as it was thought that personal knowledge of the channel’s explicit subject-matter would facilitate interpretation of the associated diagrams.

most part, this simply specifies that the *Kubuntu version* at the time was *12.10*. This cluster is, with no surprise, adjacent to *cluster 01*, while being located slightly more at the periphery of the diagram, perhaps denoting the increasing technical specificity of most terms involved and thus their comparatively distinctive patterns of use. The conceptual proximity of *cluster 06* and *cluster 01* is further confirmed by the respectively extremely-strong and strong word-associations “kubuntu—12.10” ( $C = 0,08$ ) and “kubuntu—version” ( $C = 0,06$ ). The inclusion of the word-stem “mean” in *cluster 06*, however, is more difficult to interpret. In fact, this word-stem is positioned quite far from the rest of the cluster, and thus probably has a pattern of use that overlaps only relatively slightly with *English-language Ubuntu-related channels* that of the other signs in the cluster. As for previous Ubuntu-related diagrams, it is nonetheless reasonable to think that in a channel where interactions are centred around problem-solving, some participants might wonder what some technical jargon (most often about an issue) encountered *means*, or others may request more clarity regarding the exact *meaning* of an issue someone is trying to describe, in order to better provide assistance.

*Cluster 04* includes the word-stems “BluesKaj”, “thanks”, “please”, “question”, and “help”, and similarly to analogous clusters in the preceding two sets of diagrams, indicates that people join #kubuntu primarily to ask *questions* about the operating system and seek *help* resolving issues pertaining to its use. The sign “BluesKaj” seems to refer to the name of an important contributor to the channel, which other participants often *thank* – most likely for offering solutions to their problems – as highlighted by its proximity to the word-stem “thanks”.

In the corresponding “Co-Occurrence Network of Words”, the signs “please”, “help”, and “question” are very strongly interconnected as “please—help” ( $C = 0,13$ ), the strongest association in

Figure 5b, and moderately so as “please—question” ( $C = 0,03$ ). “BluesKaj” is equally moderately associated with “thanks” and “question”, as the user to which it refers is probably often solicited for *questions*, and as mentioned above, subsequently *thanked* for solutions.

*Cluster 03* groups together “problem”, “hello”, “remove”, “driver”, “update”, and “error”. The inclusion of the word-stem “hello” in the same cluster as “problem” (and similarly close to “help” in *cluster 04*) indicates that channel users might first introduce themselves, and immediately follow by signalling that they have a *problem* they require *help* with. This highlights, once more, that users indeed tend to join the #kubuntu channel to seek assistance for *problems* they experience with the operating system. Furthermore, the presence of the word-stem “driver” seems to indicate that a considerable part of the *problems* (or *errors*) is related to *drivers*, which might require something to be *updated* or *removed*. The extremely strong associations “install—driver” ( $C = 0,09$ ) and “driver—issue” ( $C = 0,08$ ) reinforces the idea that users might experience *issues* with *drivers* that have recently been *installed*, or should *install* some *drivers* to resolve *issues*.

*Cluster 09* containing “Ubuntu”, “issue”, “KDE”, and “application”, as well as *cluster 05* which includes “screen”, “set”, “setting”, “kde”, and “window”, both highlight once again the particular aspects of Kubuntu that users tend to have *issues* with. Namely, the *KDE* (or *kde*) desktop environment, including the *applications* it encompasses, the *screen* features users can *set*, and the general *settings*. The word-stem “window” is included in *cluster 05* most likely because of its proximity to “setting”, which could either mean that users tend to mention *KDE window settings*, or the *settings* of the Windows operating system, the former interpretation being slightly more plausible in relation with the rest of the included word-stems.

Furthermore, the extremely strong word-association “KDE—version” ( $C = 0,08$ ) and the strong one “application—version” ( $C = 0,06$ ), in [Figure 5b](#), denote that *cluster 09* comprises signs that also follow a pattern of use similar to the ones in *cluster 06*. Indeed, the *issues* users experience with the *KDE* desktop environment most likely concern its most recent *version* – or, possibly, pertain to the version of some KDE software being inadequate, such as when critical updates are required.

*Cluster 02* combines the word-stems “system”, “boot”, “linux”, “partition”, “file”, “drive”, and “user”, which are all strictly technical in nature in this context. Let us take as starting point the distinctive sign “linux”, which as mentioned in a previous section<sup>131</sup>, can either be used in the context of its specific meaning as the core part of the system – the kernel – or more broadly as a generic way of referring to the operating system as a whole, especially when “*Linux*” is contrasted with the *Windows* operating system, as shown by its proximity to “window” in *cluster 05*. The fact that “linux” and “window” are both close to “boot” and “partition” – itself next to “drive”, as one would expect given that *partitions* are used to divide hard disk *drives* – is likely to mirror such technical matters as the fact that different partitions are necessary to dual-*boot* “Linux” and Windows, which is further reflected in the strongly associated pair “boot—window” ( $C = 0,06$ ). Otherwise, of course, “boot” refers first and foremost to the process of turning on a computer system.

An additional conceptual link between *cluster 02* and *cluster 05* is characterized by the second strongest association in [Figure 5b](#), “system—setting” ( $C = 0,12$ ), which should be self-explanatory in light of previous paragraphs. Two more word-stems included in *cluster 02* have yet to be made sense of: “file” and “user”. The former is simply the typical way for modern operating

131 In section [4.3 \(a\) – Interpretation of Figure 3a and Figure 3b](#).

systems to divide and label information, on a most fundamental level by making use of a certain way of structuring and retrieving data called a *file system*<sup>132</sup>, an expression mirrored by the very strong word-association “file—system” ( $C = 0,07$ ). As for the latter, its primary use here is the same as previously explained: setting up multiple *users* on a single instance of an operating system enables different people to use different settings and keep their data separate from each other’s.

*Cluster 08*, like the corresponding cluster in the previous section’s [Figure 5a](#), includes “sudo” and “command” towards the periphery of the diagram. Again, the former is commonly used to begin a command on Ubuntu and Ubuntu-based systems (in this case, Kubuntu). For instance, the strong word-association “sudo—update” ( $C = 0,06$ ) in the corresponding “Co-Occurrence Network of Words” probably refers to the routine command below, used to ensure that the operating system remains up-to-date, the importance of which is also echoed by the very strong word-association “system—update” ( $C = 0,07$ ) (more specifically, this command replaces the list of available software, on one’s computer, by the most recent information available):

```
sudo apt update
```

As for the word-stem “add”, the only other sign contained in *cluster 08*, its shared location with “sudo” and “command” could indicate that channel users might typically suggest to *add* elements to the system, by using *commands* – often requiring administrative privileges in the case of “sudo” – to fix some problems. Alternatively, being in a cluster with “sudo”, “add” might also – though somewhat less likely – be used as part of some commonly executed commands on Kubuntu,

132 See the Wikipedia page: [https://en.wikipedia.org/wiki/File\\_system](https://en.wikipedia.org/wiki/File_system).

133 In nearly all contexts of “normal” use, this command is functionally equivalent with the older form:  
sudo apt-get update

such as the following example, to add the third-party software repository *Kubuntu Backports* to one's operating system:

```
sudo add-apt-repository ppa:kubuntu-ppa/backports
```

Again, in [Figure 5a](#), “sudo” is located particularly far from other signs as it is unlikely to ever be used as part of ordinary speech. On the other hand, although “add” and “command” are also technical terms, they can sometimes be included within casual sentences, and thus figure closer to other elements in the diagram.

#### 4. 3 (d) – Interpretation of *Figure 6a* and *Figure 6b*:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“lol”, “frenocha”, “people”, “phunyguy”, “try”, “rww”, “johnjohn101”, “mean”, “run”, “window”,  
“seednode”, “roasted”, “install”, “buy”, “ubuntu”

The beginning of this list is reminiscent of the *#libertyinmind* (2013-10-29 to 2014-08-04) and *#politics* (2012-04-29 to 2014-08-04) data-sets, where the three most frequent signs were also, regardless of the exact sequence, “people” and “lol” along with a pseudonym (“Aerides”, instead of “frenocha” which appears here). We also find “try” and “mean” at a position in the list most similar to the one they occupied among the most frequent word-stems in the two other data-sets just mentioned.

The rest of the list, however, while it includes some word-stems also found among the most frequent ones in other Ubuntu-related data-sets (“run”, “install”, “ubuntu” in all of them and “window” in *#kubuntu* (2013-01-01 to 2013-01-31)), is otherwise unique in that it includes more pseudonyms

than any other such list in all the data-sets studied. In [Figure 6a](#), most of these pseudonyms (“AtomicSpark”, “blitz”, “johnjohn101”, “samthewildone”, “seednode”, “facepalm”, “roasted”, “not\_roasted”) are positioned towards the periphery of the diagram. The only exceptions, “frenocha”, “phunyguy”<sup>134</sup>, and “rww”<sup>135</sup>, also happen to be the most frequently occurring (which entails that they are more likely to occur across a greater variety of contexts), in addition to the last two – who happen to have operator privileges in the channel! – being also relatively central in [Figure 6b](#), while “frenocha” is altogether excluded from that diagram.

Many of the individual clusters in [Figure 6a](#) are more difficult to interpret than most clusters in every other “Multi-Dimensional Scaling” diagram analyzed until now. Moreover, as illustrated by the unusual absence of a significant well-defined central cluster, the structure of [Figure 6a](#) is abnormally fragmented for a diagram of that type. However, both its opaque meaning and its fragmentation are representative of the casual socialization and highly personal conversations lacking a single theme<sup>136</sup> that typically take place on #ubuntu-offtopic. Similarly, in [Figure 6b](#), the fact that so many words are “central” represents the lack of a subject-matter common to most conversations.

Despite the arcane nature of some clusters – for instance, those including only one pseudonym and one word-stem: *cluster 06* (“blitz” and “eat”), *cluster 09* (“facepalm” and “wait”), and *cluster 11* (“AtomicSpark” and “read”) – it is possible to satisfactorily decipher many more.

134 Rob Tongue, who according to his personal page on the Ubuntu wiki, last edited in 2014, was at the time “an operator in the core [Ubuntu-related IRC] channels, with focus on #ubuntu, #ubuntu-offtopic, and #ubuntu-server” (emphasis mine). [<https://wiki.ubuntu.com/phunyguy>].

135 Also an IRC operator in the following channels: “#ubuntu, #ubuntu+1, #ubuntu-devel, #ubuntu-offtopic, #ubuntu-ops, and #ubuntu-women” (emphasis mine). [<https://wiki.ubuntu.com/rww>]

136 The lack of a homogeneous subject-matter is consistent with the definition of “off-topic” – which according to Google is “(of a comment) not relevant to the subject under discussion” [<https://www.google.com/search?q=define+off-topic>] – if we assume that many, if not most, conversations in the channel are characterized merely by their relative lack of relevance to what is otherwise the subject of Ubuntu-related channels.

On that account, we may first take a closer look at *cluster 10*, which is arguably the most meaningful with respect to the distinctive purpose that determines most participants' actions in that particular channel. Apart from the pseudonym "samthewildone" at the periphery, which will be ignored for this part of the current analysis, this cluster includes the elements "play", "game", "watch", and "love". We might reasonably assume that the channel's participants regularly engage in discussions related in one way or another to how they *love to play games* or *watch* others doing so. This interpretation is substantiated by the word-association "play—game" ( $C = 0,07$ ) being the strongest in [Figure 6b](#), while "game" is one of the most central word-stems in that graph, and "love" as well as "watch" are also relatively central.

The specially close-knit *cluster 04*, by contrast with the previous one, represents what #ubuntu-offtopic has in common with other Ubuntu-related channels. Thus, it consists exclusively of terms with a specific meaning in the context of computing, which have been explained in detail in previous sections: "install", "run", "linux", "window", "ubuntu", "server", and "system". Their shared technical nature implies similar patterns of use compared to nearly all other elements in the diagram. Among *cluster 04*'s elements, "ubuntu", "run", and "linux" (in approximate decreasing order of centrality) are all some of the most central words in [Figure 6b](#), with "window"<sup>137</sup> and "install" also being somewhat central. The contexts in which these words tend to be used is further illustrated by the very strong word-associations "install—ubuntu" ( $C = 0,04$ ) and "run—server" ( $C = 0,04$ ), and by the moderately strong ones "linux—window" ( $C = 0,03$ ) and "run—linux" ( $C = 0,03$ ).

137 This word-stem is, once more, most likely the result of "Windows" being erroneously shortened during the stemming process. The word probably mostly refers to the operating system bearing that name, often despite being written without a capital letter, as reflected by "window" overlapping "linux" in [Figure 6a](#).

Only a single other cluster in [Figure 6a](#), namely *cluster 07*, located very close to *cluster 04*, includes word-stems which specifically refer to computing: “UBUNTU”, “WINDOWS”, and “laptop” (the latter being also very central in the other diagram, highlighting its significance). Of course, the first two among the three symbols just mentioned simply refer to two operating systems which might be installed on a personal computer, such as one’s *laptop*.

In *cluster 01*, three of the four most common pseudonyms in the data-set, namely “phunyguy”, “frenocha”, and “johnjohn101”, are found next to “lol”, which suggests that these three people are regularly mentioned in the same context as someone makes use of the abbreviation *lol* “to draw attention to a joke or amusing statement, or to express amusement”<sup>138</sup>. On a related note, it is worth noting that a unique feature of this data-set is how “lol” is by far the most frequent word-stem, which reflects the unquestionably whimsical nature of the channel – however, it is absent from [Figure 6b](#), meaning that it is not part of any significant word-associations (indeed, it is often used alone).

Although the inclusion of “problem” in *cluster 08* remains inexplicable, the sign’s proximity to *cluster 04* examined previously – in particular, it is closest to “ubuntu” and “server” – indicates that the usual context of its use most likely concerns the same kind of technical issues as are typically encountered in the other Ubuntu-related data-sets. The two other elements in *cluster 08*, “roasted” and “not\_roasted”, make perfect sense within the same cluster as they are two pseudonyms used by the same person.

*Cluster 03*, for its part, does not appear to clearly relate to computing, either directly or indirectly. It groups together the word-stems “rww” next to “channel” – the former being one of the

138 Quoting the definition provided by the *Oxford Dictionaries*: <https://en.oxforddictionaries.com/definition/lol>.

*channel*'s operators – along with “mean”, “idea”, and “happen”. This might be interpreted as echoing in part how #ubuntu-offtopic provides a space for the convivial exchange of *ideas*, where people frequently discuss what said ideas *mean*, as well as any relevant events that have *happened* or might happen. The word-stem “happen” appears to correspond to a rather distinctive pattern of use, as it is markedly found at the periphery of [Figure 6a](#) as well as the two other “Multi-Dimensional Scaling” diagrams which include the word, [Figure 1a](#) and [Figure 2a](#) from previous sections.

Finally, *cluster 05* contains the three word-stems “seednode”, “buy”, and “phone”. While the first element listed is simply yet another pseudonym which will be disregarded for the purpose of the current inquiry, the other two elements seem likely to refer to how the users of #ubuntu-offtopic – who probably often use their *phone* when participating in the channel to pass the time – also occasionally discuss how they want to *buy* (or maybe have bought) various things such as a *phone* or a laptop. Although in [Figure 6b](#), the word-associations “buy—phone” ( $C = 0,02$ ), and “buy—laptop” ( $C = 0,02$ ) for that matter, are somewhat weak, they nevertheless provide a picture of how these words are most often used ‘in’ the channel.

#### 4. 3 (e) – Combined analysis of all Ubuntu-related diagrams:

Some of the diagrams representing Ubuntu-related channels overlap more extensively than others: those which correspond to the ‘*all Ubuntu-related channels*’ (2004-12-25) data-set and the ‘*English-language Ubuntu-related channels*’ (2014-07-01 to 2014-08-31) data-set, respectively. That is to say, the two data-sets displaying the most similar patterns happen to circumscribe very similar channels, but quite different time-frames: respectively Christmas day in 2004, and July and August of 2014. Across

a decade and seemingly disregarding the difference in the lengths of time selected, we find very similar patterns in Ubuntu-related channels *taken as a whole*. Also extremely similar are the diagrams representing the *#kubuntu* (2013-01-01 to 2013-01-31) data-set, consisting of a single channel which is also part of the other two data-sets mentioned above, but during the month of January 2013.

The 8 *most frequent* word-stems in [Figures 3a](#), [3b](#), [4a](#), and [4b](#), although they do not appear in the exact same order, are nonetheless identical: “install”, “try”, “ubuntu”, “file”, “package”, “run”, “thanks”, “problem”. The same is nearly true of [Figures 5a](#) and [5b](#), albeit with “kubuntu” instead of “ubuntu” (which have essentially the same use in this context), and “system” instead of “thanks”.

In [Figures 3b](#), [4b](#), and [5b](#), the word-stem “install” is either the most central or the second most central, while “install—ubuntu” (or “install—kubuntu”) and “install—package” are among the strongest word-associations. This is partly mirrored, in the “Multi-Dimensional Scaling” diagrams [Figures 3a](#), [4a](#), and [5a](#), by how “ubuntu” (or “kubuntu”) is consistently found right next to “install”, either in the same cluster ([Figure 3a](#)) or in a different cluster ([Figure 4a](#) and [Figure 5a](#)). However, perhaps most remarkably, in the same three diagrams “install”, “try”, and “run” are included in the central cluster in all three cases. Still in the centre of these diagrams, “problem” is closest to “try” and “install”, albeit in a different cluster, which nevertheless indicates that people very frequently have *problems* with the *installation* of software – including, of course, the Ubuntu (or Kubuntu) operating system itself – and that a necessary part of solving such problems is to *try* possible solutions.

In addition, “please—help” is the strongest association in [Figure 5b](#), among the strongest in [Figure 3b](#), and also moderately strong in [Figure 4b](#). Similarly, in [Figures 4a](#) and [5a](#), “please”, “help”, and “question” are found next to each other within the same cluster, located away from the centre

relative to other elements – we also find “please” and “help” relatively close to each other and similarly located in [Figure 3a](#), but within two different clusters.

Another striking similarity is how the elements “sudo” and “command” are in the same cluster and located towards the periphery in [Figures 3a](#), [4a](#), and [5a](#). This is also true of the word-stems “right” and “bug” included in [Figures 3a](#) and [4a](#), in addition to the word-stem “file”, included in the central cluster in the same two diagrams. Similarly, we find the sign “thanks” in analogous locations at the periphery in all three diagrams.

It is also worth noting that the remarkable structural similarity between the ‘*all Ubuntu-related channels*’<sub>(2004-12-25)</sub> data-set, the ‘*English-language Ubuntu-related channels*’<sub>(2014-07-01 to 2014-08-31)</sub> data-set, and the *#kubuntu*<sub>(2013-01-01 to 2013-01-31)</sub> data-set is further highlighted by [Figures III](#), [IV](#), and [V](#), where the word-stems “install” and “use” consistently appear as specially central word-stems which are also systematically the two most frequent ones.

In short, the *final cause* (or purpose) of the vast majority of users' presence on such channels consists in their volition to either obtain or impart technical knowledge related to the Ubuntu operating system and its derivatives as well as any software an Ubuntu user might install. Indeed, with the exception of such word-stems as “please”, “help”, “problem”, and “thanks”, other similarities between these data-sets consists of technical terminology related to computing. That apparent purpose appears unfazed by even the most culturally important events of off-line life, as even diagrams representing data from Christmas day – the ‘*all Ubuntu-related channels*’<sub>(2004-12-25)</sub> data-set – do not include a single term associated with Christmas.

Interestingly, the *#ubuntu-offtopic* channel seems to display a combination of some patterns

otherwise found, on the one hand, in the more convivial channels #libertyinmind and #politics, and on the other hand, in the generally more impersonal and functionally-oriented Ubuntu-related channel. This hybrid nature can perhaps best be understood at a glance by comparing, in the relevant appendix, Figure VI with the two groups of diagrams preceding it – Figures I and II by contrast with Figures III, IV, and V. Indeed, both “people—think” and “people—use” word-associations are important parts of this diagram, while “people” and “use” are the two most central words.

Otherwise, the word-stem “install” is neither especially frequent nor central in Figures 6a and 6b representing *#ubuntu-offtopic* (2014-08-08 to 2014-10-13), unlike in diagrams representing the other three data-sets – this accurately represents how #ubuntu-offtopic, unlike most channels related to the Ubuntu project, is not primarily meant for its users to seek or impart technical assistance, such as what would require users to *install* software. Yet, Figure 6a does include “install” within *cluster 04*, which along with *cluster 07* groups together all the computing jargon included in the diagram. It seems reasonable to infer that although most conversations in the channel, albeit heterogeneous, are dedicated to playful socializing, a minority of the conversations revolve around Ubuntu and associated topics within the realm of computing.

Again, it is worth noting that the similarities observed within Ubuntu-related channels reflect stable patterns of sign use within communities possessing shared interests (here, the community of users of the Ubuntu operating system and derivatives), *despite the absence of a stable and explicitly defined topic*. Otherwise, the shared characteristics within the #ubuntu-offtopic, #libertyinmind and #politics channels highlight the persistence of some patterns of sign use associated with non-technical channels dedicated to more casual conversation.

## **5 – Conclusions**

### **5. 1 – Findings of Prior Research that Are Corroborated by this Work**

#### **5. 1 (a) – The structure of IRC conversations mirrors the platform's technical attributes:**

A reasonably well-established fact in communication research is that every communication tool – in spite of the inevitable variability of actual uses – *determines in some respects the form taken by communication mediated through it*. This should not be taken to be advocating the (much-maligned) theory of so-called “technological determinism”, but instead should be understood as a more trivial sort of observation according to which, for instance (to list two of the most self-evident examples I could think of): on-line platforms such as IRC or chat preclude the direct inclusion of sound, images, or video into conversations; similarly, a platform such as SMS precludes the simultaneous participation of multiple people within a single conversation.

We may then turn our attention towards some of the most salient and distinctive features of the form of communication made possible by IRC technology. However, any concise description of the platform, for example as “a versatile digital technology primarily designed, and generally used, for text-based synchronous group communication via the Internet”, does not adequately reflect how the IRC protocol is an extremely flexible one:

a) the conversations can, and often do, involve large numbers of people at once, but this is not a given<sup>139</sup>;

b) communication may be either synchronous or asynchronous<sup>140</sup>;

139 At one extreme, IRC supports “private messages” from an individual user to another, but at the other extreme, there is no limit to the number of users who may be simultaneously present or active in a single channel.

140 Although it is customary to converse in real-time, it is also possible and not uncommon for a participant to

- c) the people involved may know each other anywhere from intimately to not at all<sup>141</sup>;
- d) the conversations may be private and transient, public and permanently recorded, or in-between<sup>142</sup>;
- e) the level of prior planning and moderation may range from as spontaneous as is feasible on-line to rigidly organized<sup>143</sup>.

Perhaps more than anything else, the list above highlights how communication via IRC may take a great variety of shapes. Only a subset of the possible forms were analyzed in the present study: publicly available channels with a minimum of several different active users – thus excluding some of the possibilities described in points a) and d) above. Nevertheless, the channels selected do illustrate, on the one hand, some of the (many) characteristics that IRC has in common with other forms of communication, and on the other hand, some of the (few) characteristics that may be unique to IRC.

Doell [2000] concluded, in explicit agreement with prior research, that IRC interactions are structured so as to be more similar to “casual spoken conversation” as opposed to “any traditional written genres” [*Ibid.*, p. ix]. Likewise, his conclusions “supported the findings of earlier studies on synchronous [computer-mediated communication]” [*Ibid.*, p. 136], implying that IRC conversations are structurally analogous to other forms of real-time, text-based on-line communication.

Although the present study did not include a systematic quantitative examination of IRC in

respond much later to a message written previously – thus, it is inaccurate to describe IRC as merely synchronous, as some researchers have erroneously done [e.g. Reid, 1991].

141 While some users decide to keep using the same pseudonym for very long periods of time, nearly all public IRC networks, by default, *require* nothing more than a temporary pseudonym as “personal identification” from users. Any additional registration is completely optional and doing so usually yields only minor benefits, the most obvious being to prevent others from “stealing” one’s chosen “pseudonym” or “nickname” while one is off-line.

142 Every participant’s IRC client may be configured to keep any proportion of the content of channels joined and of private messages sent and received. It is also relatively easy and common to use a “bot” to automate the process of recording a channel, usually to later publish the resulting text data on-line

143 It is initially up to the creator of a channel to give administrative privileges to other users, after which all users with such privileges are responsible for setting the channel’s topic and enforcing its rules, the latter notably by “kicking” and “banning” those who break said rules.

contrast with other media, its results indicate that Doell's characterization of the structure of IRC discourse [*Ibid.*], while true in most cases, is incomplete. Indeed, the simple observation of a normal IRC conversation is sufficient to confirm that the attributes listed above are more or less the same as those that make possible the great variety of forms that face-to-face conversations may take – with the exception of point b), which is unique to recorded communication.

Among the features of IRC communication that may shed light on how it compares to face-to-face communication, perhaps the most notable is the fact that above all, an IRC channel is defined by its permanent name, which is necessarily known by everyone who joins the channel and stands for something its users have in common – closely related to their *shared purpose* as a discourse community, and to the *frame* (in the sense given to this notion by Goffman) within which they interpret the conversations taking place there. If only because of the channel's name, the *definition of the situation* that characterizes an IRC channel is partly *explicit*, and we may thus expect that IRC conversations are most similar to face-to-face group interactions that also involve an explicit purpose.

In the case of IRC, however, this purpose may be made more or less explicit by the use of the “topic” function, which allows the channel's administrators (and possibly other users depending on the channel's settings) to add a written description to the channel – this description is more or less impermanent from one channel to another, and potentially from one time to another within the same channel, given how it may be modified at any time. It is worth noting that the relationship which exists between a channel's name and its “topic” at any given moment is highly variable, and has not been investigated in the present study.

### 5. 1 (b) – The statistical analysis of IRC logs can reveal otherwise elusive characteristics:

Content analysis is meant to study what is not directly observable<sup>144</sup>. As was hinted to in the previous section, beyond the uni-dimensional and static label in which consists the name of an IRC channel, shared purpose results from a dynamic process of collective interpretation. As social interactions intensify and complexify through computer-mediated communication, making sense of the sometimes overwhelming flow of on-line communication through first-hand qualitative impressions becomes, at best, impractical. In response, recent studies of on-line communication have often adopted statistical analysis methods in order to uncover overreaching trends, which might otherwise remain elusive.

For instance, Doell [2000]<sup>145</sup> concluded from his comparative examination of the relative frequencies at which the most common general-purpose English words (e.g. “I”, “you”, “be”, etc.) occur on IRC, as opposed to other forms of communication, that IRC conversations are most similar to casual spoken conversations rather than typical forms of written discourses.

Michels [2012]<sup>146</sup> also analyzed IRC logs (including that of the #ubuntu channel) by employing statistical methods, which essentially consisted in writing and using a computer program divided into different modules, one of which implemented a keyword matching algorithm, and another a synonym-based topic analysis... leading him to the assessment that the topic analysis results “were not all that spectacular” [Ibid., p. 52], an outcome he ascribed mainly to limitations of the stemmer software and to how the thesaurus used could not adequately handle Internet jargon.

144 As explained in section 8. 8 (a) – Content analysis & computational linguistics.

145 This study was previously summarized and commented on in section 2. 5 – Two Examples of Statistical Methods Being Used to Analyze IRC Logs.

146 Same as above.

Another such example is a study by Kramer et al. [2004], pertaining to an IRC channel where support was provided to people living with bipolar disorder. Using predefined categories of words along with relatively straightforward computational tools – ultimately resulting in simple lists of frequencies and correlations – their study nonetheless demonstrated that such quantitative methods are capable of “[making] comparisons across different online forums”.

Notwithstanding the computational benefits of quantitative analysis for its efficiency and reproducibility, the studies mentioned above lack adequate theoretical and empirical tools to provide meaningful conclusions related to the overreaching numerical trends brought to light in the data.

Along those lines, Holmer [2008] proposed to look at statistical analysis of online data in a different way, stating that “The structural properties of online discourse can be used to analyze underlying communication behaviour and social structure”. In this article, he describes and demonstrates a method, called *Discourse Structure Analysis* (DSA), that automatically analyzes chat transcripts and produces visualizations of their interactive structure.

Most importantly, this article emphasizes the usefulness of data visualization, which the author says makes “the dynamics of communication become easier to grasp”, as well as “supports subsequent qualitative analysis”. This highlights the value of incorporating quantitative methods to allow a more systematic and comprehensive qualitative account of meaningful aspects of the data. However, in the last analysis, this author solely focuses on an analysis of the dynamics of interaction among users, and does not make use of his novel computational method to reveal the meaning of the actual content of chat logs.

Nonetheless, he concludes by asserting that “A topic for future research is the usefulness of

these visualizations [...] for understanding the content of a chat log”, which adequately highlights the relevance of work such as the present thesis.

## **5. 2 – Four Hypotheses Herein Put to the Test and Substantiated by Newfound Evidence**

### **5. 2 (a) – Each form of semiosis has by-products which are isomorphic to a corresponding type of semiotic network:**

According to the epistemological presuppositions that guided this work<sup>147</sup>, the tenets of Peircean semiotic are to be understood as provisional assumptions, held to be probably true only as long as their pragmatic implications are consistently shown to be useful with respect to what they purport to accomplish. While Peirce put forward that cognition/communication is a process which occurs by way of signs and inferences that constitute the activity of signs, this hypothesis is to be understood from the angle of the falsifiable predictions it entails.

Arguably, the most crucial prediction of Peircean semiotic – as conceptualized for the purposes of this research – is that signification, as a virtual semiotic network which develops logically, is instantiated predictably along with isomorphic effects. That is to say, an agent's *habits of thinking*, insofar as they are *expressed*, should correspond systematically to observable and measurable regularities in sign-associations (hereafter termed *semiotic patterns*). Furthermore, the fundamental structure of these semiotic patterns should make it possible to represent them as complex “network-like” diagrams<sup>148</sup>.

147 Outlined in 3. 1 – Epistemological and Methodological Framework.

148 Any diagram including essentially non-hierarchical, decentralized, multidimensional relationships between multiple elements at once relative to each other. Sufficiently complex graphs generally correspond to this definition, for example, but so do other types of diagrams, like those of the “Multi-Dimensional Scaling” variety

With regard to the present endeavour, this broad presupposition cannot be stressed enough: *symbols, such as language-signs, are rules; they exhibit law-like behaviour*. They essentially originate outside us, and we use them in regular ways, according to the practical effects that we seek to produce in the interplay of the symbols used and the interpretative processes upon which they act. However, it is the symbols themselves (like all signs) which hold the ultimate determination of the process – “[...] semiosis processes [are] autonomous or self-governing<sup>149</sup>” [Ransdell, 1992, p. 3]. Building upon Peirce's insight: “just as we say that a body is in motion, and not that motion is in a body, we ought to say that we are in thought, and not that thoughts are in us” [CP 5.289 Fn P1] – the conceptual machineries of language-games are something that *we are into*, not something that *is in us*. This surely precludes neither creativity nor liberty, but it does entail that purposed, voluntary cognition/communication can achieve what it purports to do only insofar as, to some extent, it operates predictably.

The autonomy of semiosis, along with its predictability, entail that every semiosis process relates to other semiosis processes in a predictable and measurable way. In addition, the pragmatic nature of the meaning of symbols (according to which *their meaning is their use*) entails that there should exist an observable relationship between the use of symbols in a given context and the corresponding purpose. Therefore, by representing a cognition/communication process – such as a conversation – as a *network of semiotic patterns*, it should be possible to observe how that process relates to its purpose, and to measure how it compares to other similar processes.

Although the present research was directly concerned with IRC alone as a case study, and did

shown in this thesis.

149 We may add that they are *autopoeitic*, that is, self-reproducing and self-maintaining [See Lévy, 2011].

not include a systematic comparison with other means of communication, one of its aims has been to argue that the methods employed are fundamentally applicable to all forms of semiosis. More generally, I have argued that this is the case for the associated methodological paradigm, which I have briefly suggested be termed *computational semiotic*<sup>150</sup>, defined notably by largely disregarding grammatical features (in the ordinary linguistic usage of the term) of the text data for the purposes of analysis, in favour of an interpretation in light of conventional sign-associations (or “grammar” in the Wittgensteinian sense).

That is to say, albeit with very limited time and resources, I have tried to show that the use of computational tools to generate diagrammatic representations of notable statistical features can potentially be used for the scientific analysis of any instance of semiosis, so long as an observable counterpart to that process is found which may be appropriately recorded as digital data.

It was also my intent to offer evidence that such computational and diagrammatic methods may offer some substantial benefits for the scientific understanding of signification, primarily with respect to circumscribing and comparing the particular characteristics of the natural processes that generate it. From that point of view, my analysis of IRC ought to be construed as providing exploratory insight into the defining features of different instances of semiosis, so as to better understand how they relate to each other and to their context.

Among the findings of this study that have provided evidence for the notion of semiosis championed herein, we notably find the fact that all diagrams representing instances of human collective semiosis not only highlight what is manifestly the shared purpose of participants, but also

150 See [footnote #66](#) in section [3.2 – Tools Used](#).

display a similar structure by contrast to diagrams representing language generated by “bots” (an example of what Peirce called “quasi-minds”).

In addition to the nature of language as a network of signs, the corresponding nature of *society* as a network of signs was reflected within the diagrams by their simultaneous inclusion of pseudonyms as well as other language-signs, thus showing that the way in which people relate to each other – mediated by the symbolic representations they create of themselves and others – is analogous in many respects to the way in which all concepts relate to each other in a given context.

5. 2 (b) – Diagrams can reveal more aspects of complex semiotic patterns than other means to that end:

Diagrammatic representations can make readily intelligible a variety of complex logical and mathematical patterns of sign use – in the specific context of this exploratory case study, patterns occurring within and across IRC logs. Thus, the facts revealed by diagrams representing IRC logs can be described as the semiotic patterns which manifest when “semiotic rules of conduct” (roughly equivalent to Wittgenstein's notion of “grammar”) become the determination<sup>151</sup> for one's sign-generating behaviour upon each particular occasion.

As a corollary of the special epistemological role bestowed upon predominantly iconic representations by Peirce's theory of signs, the patterns revealed by diagrams are predicted to be normally more representative of their object than those revealed through other (non-diagrammatic) means of achieving similar ends.

<sup>151</sup> This occurs through the mediation of a “resolution”, defined as “a mental formula always more or less general [and] nothing more than an idea” [CP 1.592] which “is of the nature of a plan; or, as one might almost say, a diagram” [Ibid.], and is subsequently “converted into a determination” [Ibid.].

One way of verifying whether or not this is objectively true is to assess whether or not the diagrams used in the present study reveal – or could potentially reveal – *all* the types of patterns that would normally be laid bare by a more traditional statistical analysis, *in addition to* at least one type of pattern (and preferably many) that *cannot* be revealed through other methods. Indeed, the diagrams representing the IRC channels analyzed<sup>152</sup> clearly demonstrate not only the “themes” or “topics” of the conversations, but also many aspects of their structure that cannot be uncovered by simple word frequencies and distributions of frequencies (even when they are used along with a thesaurus to combine words with similar meanings [Michels, 2012], for instance).

Generally speaking, the diagrams used herein make it possible to get a sense of the context in which words are used, something which would otherwise require the researcher to look at sentences in the text data one by one. Among the many features which enable this, two of the most useful are the *word-associations* shown in “Co-Occurrence Networks of Words” diagrams, and the *clusters* shown in “Multi-Dimensional Scaling” diagrams.

Most importantly, the diagrams studied in this work also show *how multiple characteristics overlap with each other at once*, something that only diagrams can represent in a way that is understandable for a human being. Such overlapping characteristics, for example *frequency* coupled with *centrality*, have proven to be often the best way to group together similar channels.

To recapitulate, I hold that cognition/communication – whether “individual” or “collective” – is a teleological process, taking place essentially through signs understood as a system of interwoven

<sup>152</sup> Let us reiterate that even a diagram representing an instance of semiosis as it actually unfolded over a definite period of time in the past may be said to hold meaning of an essentially virtual nature, even though the elements of semiosis (inferences or logical sign-associations) depicted have in fact already been actualized. The meaning of such a diagram, in the last analysis, does not lie within it, but beyond it – in how it can be understood, that is, *further interpreted*.

rules for their use, and essentially involving interspersed inferences. A model of communication which corresponds to this view may be pictured or mapped as a *network of signs*; a sketch of such a network is found in the visualizations achieved with *KH Coder's* “Multi-Dimensional Scaling” and “Co-Occurrence Networks of Words” diagrams. These conclusions apply equally to a monologue (amounting to a solitary dialogue with oneself), to a dialogue between two people, and to all forms of collective conversation.

Construing a train of thought or a conversation as a network of signs ought to involve more parameters than the mere existence of the *vertices* and *edges* of the simplest graphs. The complex nuances of meaning arise through symmetrically complex habits of thinking, with “grades of strength varying from complete dissociation to inseparable association” [CP 5.477]<sup>153</sup>. Thus, while I argue that semiotic patterns can be most adequately represented as a diagrammatic “map” of their most significant formal (mathematical) patterns, it is worth remembering that the varieties of meaningful patterns that can be shown in a diagram of semiosis are *potentially endless*.

5. 2 (c) – This study's diagrams mirror how different “discourse communities”, or “language-games”, interpenetrate:

To use terminology borrowed from computer science, it can be useful to describe the structure of social groups as a *nested hierarchy* of imprecisely bounded discourse communities (or language-games) *recursively defining each other* by virtue of their interrelations, more particular communities being embedded within (or subsumed by) more general communities – the most general

153 As initially quoted in 2. 1 (b) – Peircean pragmatism.

of all being circumscribed by all instances of semiosis, which I propound might be called the *universe of signification*<sup>154</sup>.

Conceptualized as a discourse community, an IRC channel can be described as belonging first of all to the very general community delimited by the experience of human socialization, then more particularly to the community (and concomitant range of possibilities) delineated by computer-mediated socialization via the Internet, and still more specifically to the community distinguished by the use of text-based themed group conversation software, then finally to *that* particular community defined by *that* definite purport in *that* unique social, historical, and technological context.

If we recall that Wittgenstein construed *meaning* as being synonymous with *use* (in most cases), and that his concept of language-game emphasizes how the use of language in a particular context corresponds to “an activity, or of a form of life”<sup>155</sup> [PI §23], we might say that any particular language-game circumscribes a *form* or *pattern of sign-use*. Such a form or pattern encompasses interrelated ways in which a vaguely defined collection of language-signs are *similarly used*, the nature of this similarity being that of *family resemblance*, and a “family” being in turn characterized by attributes that include irremediably fuzzy boundaries and the members resembling each other in endlessly diverse potential ways. Thus, in Wittgensteinian terms, we can describe the structure of language as it is used within different communities as being coextensive with *interpenetrating*

154 This echoes Juri Lotman's notion of *semiosphere*: “Such a continuum we, by analogy with the concept of ‘biosphere’ introduced by V. I. Vernadsky, will call the ‘semiosphere’. We must, however, warn against any confusion between the term ‘noosphere’ used by V. I. Vernadsky and the concept of ‘semiosphere’ here introduced. The noosphere — is a specific stage in the development of the biosphere, a stage connected with human rational activity.” [2005, p. 206—207]

155 It is worth noting that Wittgenstein defined the speaking of language as a whole as a broad “form of life” (and thus a language-game), but for the purposes of the current analysis, the focus will be on potential subdivisions thereof.

*families of language-games*. Such a conception is a perspective different from, but similar to and compatible with, the aforementioned Peircean explanation in terms of discourse communities.

By definition, any particular language-game is never rigidly circumscribed. Furthermore, all language-games have something in common with other language-games, and it is possible *in theory* to draw a boundary anywhere so that language-games are grouped in any conceivable way. Regardless of how they are grouped, however, the degree of overlap between different language-games expresses interdependent features through which different communities are defined: shared rules prescribing the way language-signs are to be used (what Wittgenstein refers to as “grammar”); shared experiences, encoded as memories of which we are more or less fully aware, through which language-signs are understood (similar to Peirce's notion of “collateral knowledge” or to Goffman's concept of “frame”).

While *scientific inquiry* (which can itself be viewed as a particular language-game) is only one among countless *purposes* according to which language-games may be grouped, it does correspond to the purpose of the current thesis. Accordingly, language-games have herein been compared according to specifically scientific criteria, that is to say, *mathematical facts* about their actual structure have been compared after being generated according to a rigorous procedure which may be *exactly reproduced* by making use of the *same computation* described in the present work to process the *same text data*. This computation results in diagrams that may be said to represent the structure (or grammar) of particular language-games, which can consequently be assessed as more or less similar to each other.

In the last analysis, the case of IRC illustrates that fruitful dialogue presupposes the *sharing of a form of thought* (a range of cognitive possibilities, also described as a discourse community),

understood as a *family of customs regarding sign-use*. Such customs are a species of what might be termed *semiotic patterns*. Aside from the obvious necessity of a shared language, people self-organize according to common purposes coded by virtue of signs that represent them. Each community defines its own identity through imitation of, and opposition with, other communities.

Channels explicitly purporting to have similar themes (primarily by means of their name, and sometimes to a lesser extent their “topic”) empirically exhibit – sometimes strikingly – similar semiotic patterns, as represented by isomorphisms in the corresponding diagrams of sign-associations: most remarkably, the most frequent word-stems, as well as the most central, tend to be stable across channels of the same variety. The more similar the channels, the fewer unique semiotic patterns they display with regard to the other(s).

5. 2 (d) – Established semiotic patterns can withstand sweeping changes, including ones in setting or in participants:

It can be deduced from the tenets of Peircean semiotic that meaningful communication – with other people as well as with oneself, the latter being understood as the essence of rational and reflexive thought – necessarily occurs for a purpose, which serves as the logical “final cause” that ultimately determines the probabilistic inferences underlying all utterances subsequently generated.

In other words, people deliberately taking part in a conversation *necessarily do so for a reason*<sup>156</sup>, and structure their thoughts so that every symbol chosen for use within the conversation is

<sup>156</sup> This assertion might appear self-evident to the reader, but its importance for the study of communication (if not cognition in general) is so great that it is worthy of exhaustive conceptualization and analysis. Compare, for instance, with how the psychological theory of “priming” might seem to describe the ostensibly obvious notion that “similar stimuli are mutually associated in the mind”.

essentially determined (excluding, above all, the unavoidable presence of random chance<sup>157</sup>) by its logical relationship with the reason. Thus, the purpose of a conversation tends to impose strict restrictions on the words and sentences that can reasonably be used, or “make sense”, to fulfill that purpose. Given the *autonomous nature of semiosis (or grammar)*, each person's individual habits of conduct will tend to be of secondary importance by comparison to the *public rules (or conventions)* that prescribe how language-signs are to be used within a given discourse community or language-game. If one is to be understood, one's use of each language-sign must be similar to what others *habitually expect*<sup>158</sup>.

For the above theoretical considerations to shed light on the specific case of IRC, let us take as starting point the fact, discussed previously, that IRC conversations may take various shapes depending on factors such as the number of participants or how familiar they are with each other. What's more, these factors will necessarily vary both from one IRC channel to another, and within a single IRC channel over time.

Despite such potentially diverse and fluctuating conditions, IRC conversations are capable of remaining consistently and predictably meaningful over time. Indeed, some of the diagrams included in this study illustrate that the semiotic patterns found in a given channel during a given period will often be remarkably similar to those found in the same channel years later, when we might reasonably expect that *few participants, or none, have stayed the same* from one period to the other – highlighting that the fulfilling of particular purposes presupposes the objective adequacy of particular

157 Peirce aptly argued that randomness is inherently part of all aspects of reality, a philosophical notion which he termed “tychism”, defined as “the doctrine that absolute chance is a factor of the universe” [CP. 6.201]. This appears to have *probably* been empirically substantiated by the findings of quantum physics in the 20th century.

158 Even if one decides to give a language-sign an “utterly new” meaning, one will be *forced* to make the new definition explicit by making use of other language-signs in customary ways.

uses of some language-signs above others, in accordance with the *pragmatic implications* of the public rules they correspond to.

In fact, the similarity between two data-sets (*#libertyinmind* (2013-10-29 to 2014-08-04) and *#politics* (2012-04-29 to 2014-08-04)) reflecting channels that shared much of their active users and were partly recorded at the same time was shown to be less extensive than the similarity between data-sets (‘*all Ubuntu-related channels*’ (2004-12-25), ‘*English-language Ubuntu-related channels*’ (2014-07-01 to 2014-08-31), and *#kubuntu* (2013-01-01 to 2013-01-31) ) representing channels that are comparatively more similar with respect to their purpose, but much less so with respect to their users or the time-frame selected<sup>159</sup>.

This exploratory evidence *suggests* that the purpose of a conversation may logically determine its overreaching structure to a greater extent – and in more significant ways – than the participants' idiosyncratic habits. Similarly, the diagrams shown in the present study highlight how the fact that the participants tend to *obey the rules* of the language-games they take part in entails that they will often use specific language-signs in highly similar ways.

That is to say, because “a general sign or symbol is a mental connection pertaining to a habit of conduct”<sup>160</sup>, its use by different people for similar purposes will necessarily exhibit considerable regularity, which is indeed highlighted by the same elements frequently appearing in similar locations across different diagrams. This is, of course, most obvious in the case of symbols such as “thanks”, “lol”, or “hmm”, which can be assumed to have very similar and distinctive patterns of use across very different language-games, but many more examples were identified in section 4 – Analysis.

159 Some of the relevant diagrams – namely Figures 3a and 3b as opposed to Figures 4a and 4b – further illustrate that the patterns expressed in a channel during a single day can be analogous to those expressed in that channel over months, thus corroborating the notion that the semiotic patterns found exhibit remarkable consistency regardless of the time-frame selected.

160 2. 1 (c) – Semiotic & pragmatism as interdependent frames of reference, p. 34.

For all of the aforementioned phenomena to occur, the first criterion that must be fulfilled is the definite presence of a general purpose more or less explicitly shared by the overwhelming majority of users of a particular IRC channel (distinct from purposes unique to any user) over a given period – ideally a channel's entire existence. This *shared purpose* is functionally synonymous and thus coextensive with a *definition of the situation* agreed upon by all participants. This agreement, in the case of IRC at least, seems to normally be *tacit for the most part*, but never entirely so: indeed, every IRC channel is defined by a permanent name which every participant *must know* in order to join.

### **5.3 – Major Limitations of this Research**

#### **5.3 (a) – Lack of statistical representativeness:**

The specific IRC channels and time frames represented in the diagrams shown previously were chosen according to both arbitrary criteria<sup>161</sup> and systematic ones, with the latter category subdividing further into criteria that were either essentially randomized or not. Understanding that to achieve genuine statistical representativeness, *all selection criteria* ought to be of the essentially randomized variety, there was never any doubt for this author that the empirical observations included in the present thesis cannot be considered statistically representative, and therefore cannot be strictly generalized in a mathematical sense.

However, this methodological limitation is largely inconsequential considering that it was

161 Some channels were selected to facilitate interpretation, either because they had been administrated by this student researcher (#libertyinmind and #politics), or because their subject-matter was specially familiar to him (#kubuntu). Otherwise, one time-frame (December 25<sup>th</sup>) was selected to explore the question of whether or not outside events – such as Christmas – would influence the content of conversations.

never the aim of the present study to show *results* that would be applicable to *any IRC channel at any time* – nor even, for that matter, results that would invariably apply to any single IRC channel. Rather, with respect to its empirical and statistical component, the present endeavour attempted to demonstrate that a particular *method* could be fruitfully used for the analysis of any IRC channel(s), for any time frame(s), in such a way as to enable the channel(s) and time frame(s) chosen to be objectively compared to any other.

Thus, the results presented herein ought to be understood as an exploratory case study demonstrating how a particular method *might be used* for many purposes, notably to reach generalizable conclusions. Although it appears rather doubtful that a truly representative sample of “all IRC channels” could ever be obtained, it is surely possible to achieve a much greater level of statistical representativeness than what has been done in this thesis: for instance, random channels and time-frames could be selected within the corpus of Ubuntu-related channels, or that of Wikimedia-related channels.

### 5. 3 (b) – Choice of settings for KH Coder’s diagrams:

By contrast with the above, one issue encountered that was considerably more difficult to overcome, and has important implications regarding the adequacy of the methods used, is that KH Coder offers various settings which reveal different aspects of the corpus studied. In the case of “Multi-Dimensional Scaling” diagrams, any one of four methods can be chosen to compute the similarity in how words are used (“Classical”, “Kruskal”, “Sammon”, “SMACOF”), in addition to three algorithms that are available to specifically compute the distances between words (“Jaccard”,

“Cosine”, “Euclid”) – which entails a total of 12 combinations one can choose from, some of which result in vastly dissimilar diagrams.

In the case of “Co-Occurrence Networks of Words”, the same three algorithms used for distances in the previous type of diagram (“Jaccard”, “Cosine”, “Euclid”) may be used to compute the word-associations represented by the edges of the graph, in addition to two mutually exclusive measures which can be calculated in three ways each, namely “Centrality” (“Betweenness”, “Degree”, “Eigenvector”), selected by default, and “Communities” (“Betweenness”, “Random Walks”, “Modularity”) respectively – implying 18 possible combinations.

Finally, both types of diagrams require one to choose the precise number of words included, with “Co-Occurrence Networks of Words” also asking for the number of edges to be represented. Still, this summary excludes a long list of other settings relating to cosmetic aspects of the diagrams.

To complicate things further, the methods that seem to yield the most enlightening results are frequently not the same depending on the text data analyzed, but using different methods for diagrams of the same type within the same study would undermine the possibility of objectively comparing these diagrams with one another. Selecting the settings that appeared to yield the best results *on average* for the purposes of this research endeavour was far from being trivial, required weeks of trials-and-errors, and ultimately was to some extent arbitrary.

Ideally, given much more time and resources, this study could have been made substantially more exhaustive by selecting *several combinations of settings* for each type of diagram, and accordingly generating several different “Multi-Dimensional Scaling” diagrams and “Co-Occurrence Networks of Words” for each data-set. The experiments conducted with KH Coder by this student

researcher suggest that among the possible combinations of settings, some have a tendency to compensate for the shortcomings of others – in much the same way as different kinds of diagrams were found to generally complement each other.

Other limitations related to KH Coder, which are however of much lesser importance, include the fact that many characters – such as all punctuation marks – were excluded from analysis by the software (resulting in the exclusion of emoticons, for instance), or as noted by Michels [2012], the fact that the *stemmer* had some issues processing the idiosyncrasies of language used on-line – including, for instance, pseudonyms stylized without an initial capital letter – and at times reduced words where this was not appropriate<sup>162</sup>.

### 5. 3 (c) – Choice of stopwords:

Among the many unexpected obstacles that were encountered over the course of this study, one seems especially worthy of mention: the inevitable arbitrariness and potentially high impact of the choice of stopwords, namely words considered too general and/or too frequent for their inclusion in statistical analysis to reveal anything but the most general – and best-known – facts regarding the structure of the English language itself. Both of the Master's theses previously discussed that preceded the present work in using statistical methods for the study of IRC logs [Doell, 2000; Michels, 2012] ostensibly did not exclude *any stopwords at all* from their respective analyses.

While experimenting with various parameters to generate the diagrams that were to be used in this thesis, it was difficult, if not practically impossible, to decide on a specific rule or threshold to

<sup>162</sup> This was sometimes even the case for my own pseudonym, “Aerides”, which was occasionally written by other participants without capitalization and thus shortened to “aeride” by the software.

systematically exclude certain words. Five predefined lists of stopwords, and dozens of variations thereof, were thus tested, always with limited success as meaningful and unique associations involving some of the most frequent and general words in the English language were found to occur in many instances in the data-sets analyzed.

Ultimately, while the default list of stopwords provided by the GNU R software (itself used extensively by KH Coder), with a few minor adjustments<sup>163</sup>, was used for most diagrams, additional “Co-Occurrence Network of Words” diagrams are shown in *Appendix 6*<sup>164</sup> that include a number of additional verbs otherwise ignored as stopwords.

#### **5. 4 – Conceivable Practical Uses for the Findings Now Under Consideration**

In the case of complex collective conversations that emerge nearly spontaneously on-line (like they do on IRC or forums), it is often difficult to know what is being discussed with sufficient accuracy to compare a given on-line community with others. In other words, even for the participants themselves, there is seldom a reliable way to represent the shared purpose expressed through a collective conversation. A community's name provides only hints of that information, as users will inevitably interpret the meaning of their participation to a conversation in multiple ways that go far beyond any label(s) initially assigned.

As a result, it can be a daunting task for someone to navigate through the myriad of existing on-line communities to find the one(s) best suited to their purpose. While this search can be made easier by the use of any classification system, the somewhat chaotic nature, and sheer volume, of the

163 The exact list used is included in section 8. 5 – Appendix 5: List of Stopwords Used.

164 Specifically in 8. 11 – Appendix 11: Additional Diagrams – Including Additional Verbs.

content involved makes any effort to *manually* categorize on-line communities tedious and ultimately of limited value.

To remedy this problem, the theoretical framework herein set forth, and subsequently put to the test, has two relevant implications. Firstly, it entails that it is possible – such as by computational means – to generate representations of meaningful on-line content which bear a *verifiable causal relationship* with that content's underlying semiosis processes and their signification. Secondly, and most importantly, it points to diagrams as being uniquely useful for objective reasoning, given how a diagram is “[mainly] an Icon of the forms of relations in the constitution of its Object” [CP 4.531] – that is to say, the patterns in a diagram are isomorphic to some of the patterns that define what it represents.

Thus, a computational tool like KH Coder should theoretically be capable of producing diagrams that verifiably reveal defining aspects of what any on-line collective conversation signifies, no matter how complex or voluminous. Diagrams so generated would bring to light, at the very least, the predominant *implicit* subject matter of the conversations represented – in a way similar to any sufficiently complex and adequately performed statistical analysis. However, they would also show multiple other facts pertaining to the conversation, including (but far from being limited to):

- any combination of signs that is particularly distinctive or meaningful with respect to the conversation's purpose;
- the kind of context in which common signs tend to be used;
- the participants that are most central to the conversation, if any.

While many tools already exist to generate diagrams representing text data, the creation of

software specifically designed for the comparison of on-line communities would bring important benefits. For instance, such a system could make it straightforward to select any time frame, so as to understand how a particular community is shaped by the passage of time and outside events. Ideally, the diagrams could even be made to represent the fundamentally dynamic nature of conversations: the evolution of a collective conversation over time could thus be shown by generating multiple “frames” representing its content at specific moments in time separated by regular intervals.

A specially constructed tool should also provide some level of flexibility with respect to so-called stopwords, perhaps suggesting a standard list (or many standardized options) but also allowing the user to simply select or un-select words at will. This would make it possible for one's comparison of on-line communities to take into account different levels of generality/particularity with respect to the patterns under consideration.

A similar tool could be designed to facilitate the selection of keywords (or “key-expressions”) to summarize academic papers or other writings. This could be described as a “visually assisted tagging” system, as the diagrams generated would merely offer useful recommendations for the author to potentially select from.

## **5. 5 – Overview of Paths for Further Research**

Research seeking to follow into my footsteps, if it ever occurs, should take note that the epistemological axiom most crucial to my work was the *requirement of computability* as criterion for scientificity – from which can be inferred the special adequacy of methods inspired from *quantitative content analysis* and *computational linguistics* (referring to the approach employed in this research,

we might speak of *computational semiotics*) for the empirical study of semiosis in real-life circumstances, especially those involving computer-mediated communication, as they yield *computable data* by definition. It is the fact that this immense amount of data is computable that has made possible countless new ways to study human society in a genuinely scientific fashion.

The most pressing question that ought to be asked with respect to the conclusions of this thesis is to what extent they are a result of the nature of the medium studied as opposed to the nature of cognition, language, or semiosis in the most general sense. To address that question, it would be appropriate to perform an exhaustive formal comparison with a variety of other media and forms of communication (face-to-face conversations between two people, group meetings in person, text messaging, Internet forums, Facebook or Twitter, etc.) – and potentially some forms of *cognition* – through the very same hybrid quantitative—qualitative—interpretative methodology that was applied here. It is only due to constraints of time and resources that such a comparison was only marginally carried out here.

Similarly, it would also be appropriate to design a procedure to ascertain whether or not my specific observations as to the law-like behaviour of signs are in fact reducible to the laws of psychology or any other particular science – although the specific nature of what such a procedure would be eludes me. Following the philosophical framework chosen, I have generally *assumed* that regularities in sign-association can be construed as a matter of semiotic rather than psychology or biology, all the while being aware that this assumption has yet to rely on sufficient empirical grounds and must be understood as a probable hypothesis. I believe that my work has provided some evidence to that effect, but plausible contrary hypotheses ought to be explicitly tested against my own.

Another option to build upon the results of the present research would be to perform a longitudinal study either of the same or other chat-logs – ideally displaying the results as an *interactive* visualization to mirror more closely the nature of communication through IRC. Yet another possibility would be to compare the structure of the speech of different individual users (which, however, would probably raise additional ethical hurdles). Publicly available logs<sup>165</sup> lend themselves well to such analyzes as those channels have frequently been exhaustively recorded for years. Again, it is mainly due to lack of time and resources that my own analysis has only briefly touched upon the semiotic evolution of chat-rooms over time and the individual features of each user.

Finally, among the paths which seem specially worthy of consideration to this student researcher, one appears such that it could enable valuable developments regarding potential practical applications for this work's findings<sup>166</sup>.

This consists in the prospect of designing algorithms that would formalize and automate the strictly quantitative measurement of the structural similarity between diagrams displaying computational semiotics data (only to be used when the diagrams are of the same type, generated using the same settings). Such algorithms could, for instance, produce one number summarizing the degree of isomorphism between entire diagrams, in addition to many separate numbers corresponding to the discrete categories of information represented in diagrams of that type, quantifying the diagrams' level of isomorphism with respect to each of these particular aspects.

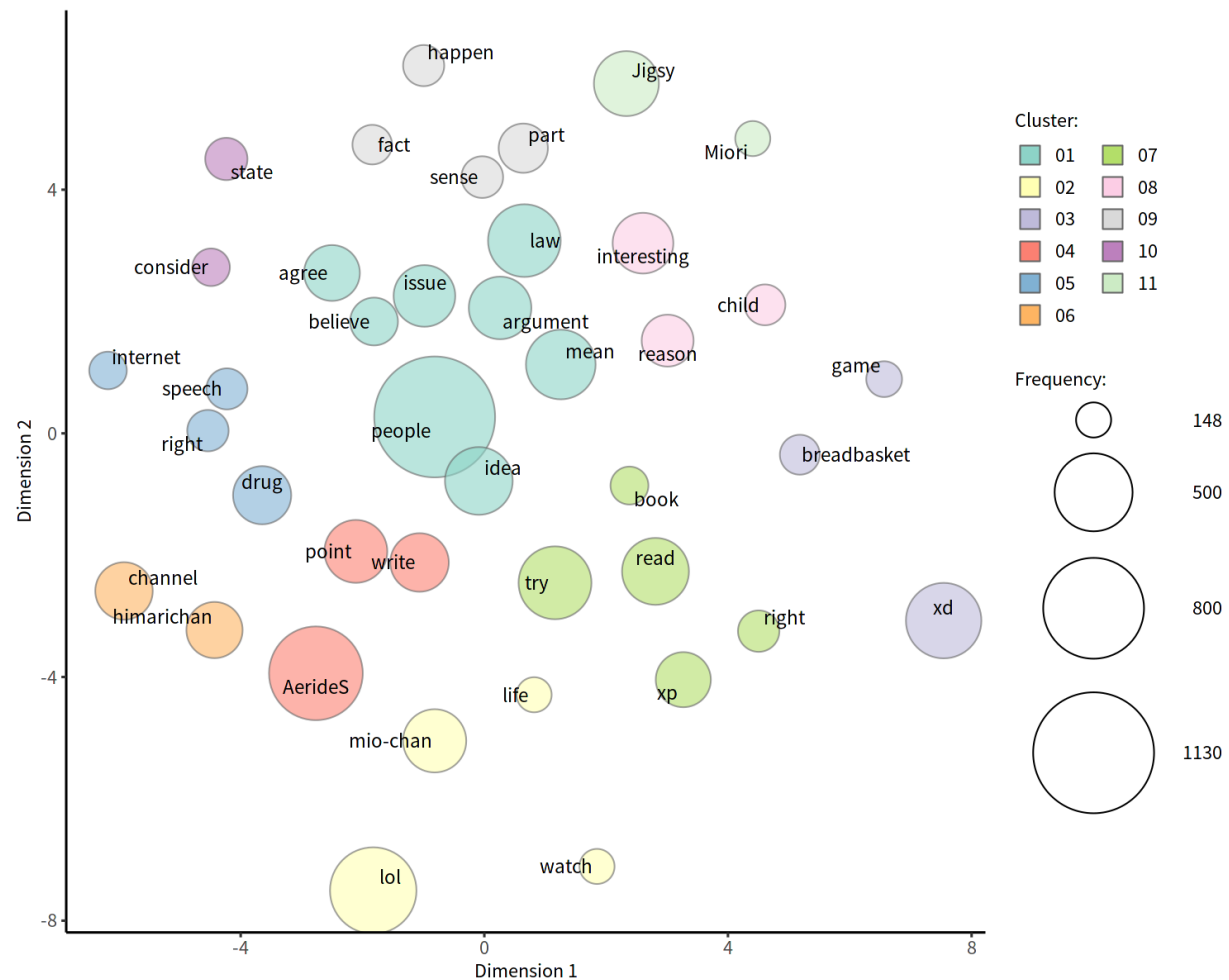
165 Analyzed in sections 4. 3 – Ubuntu-related channels (KH Coder, Part 2) and 8. 10 – Appendix 10: Additional Diagrams – Wikimedia-related channels.

166 As described in the previous section, 5. 4 – Conceivable Practical Uses for the Findings Now Under Consideration.

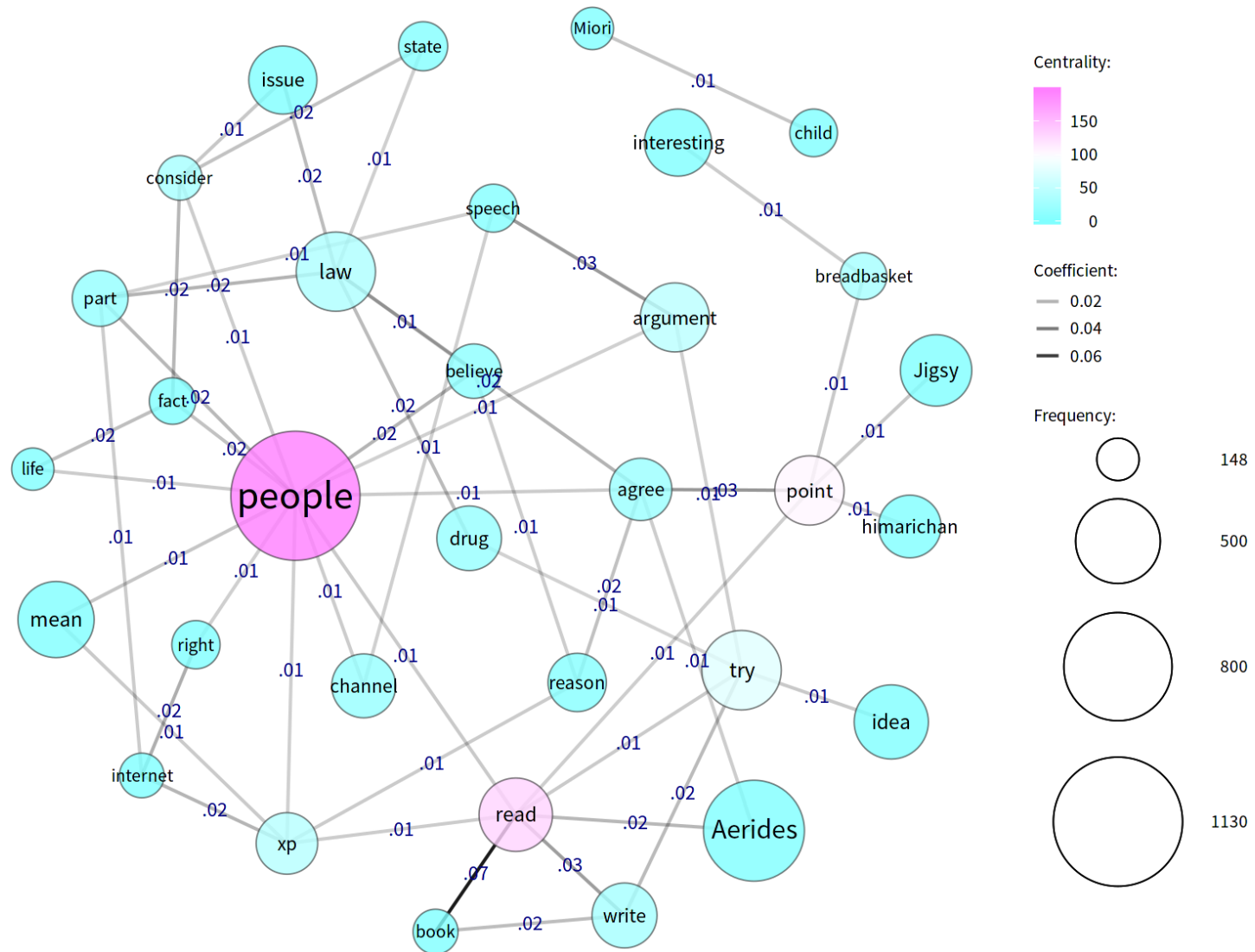
## 6 – Primary Diagrammatic Representations

### 6.1 – Diagrammatic Representations of #libertyinmind and #politics

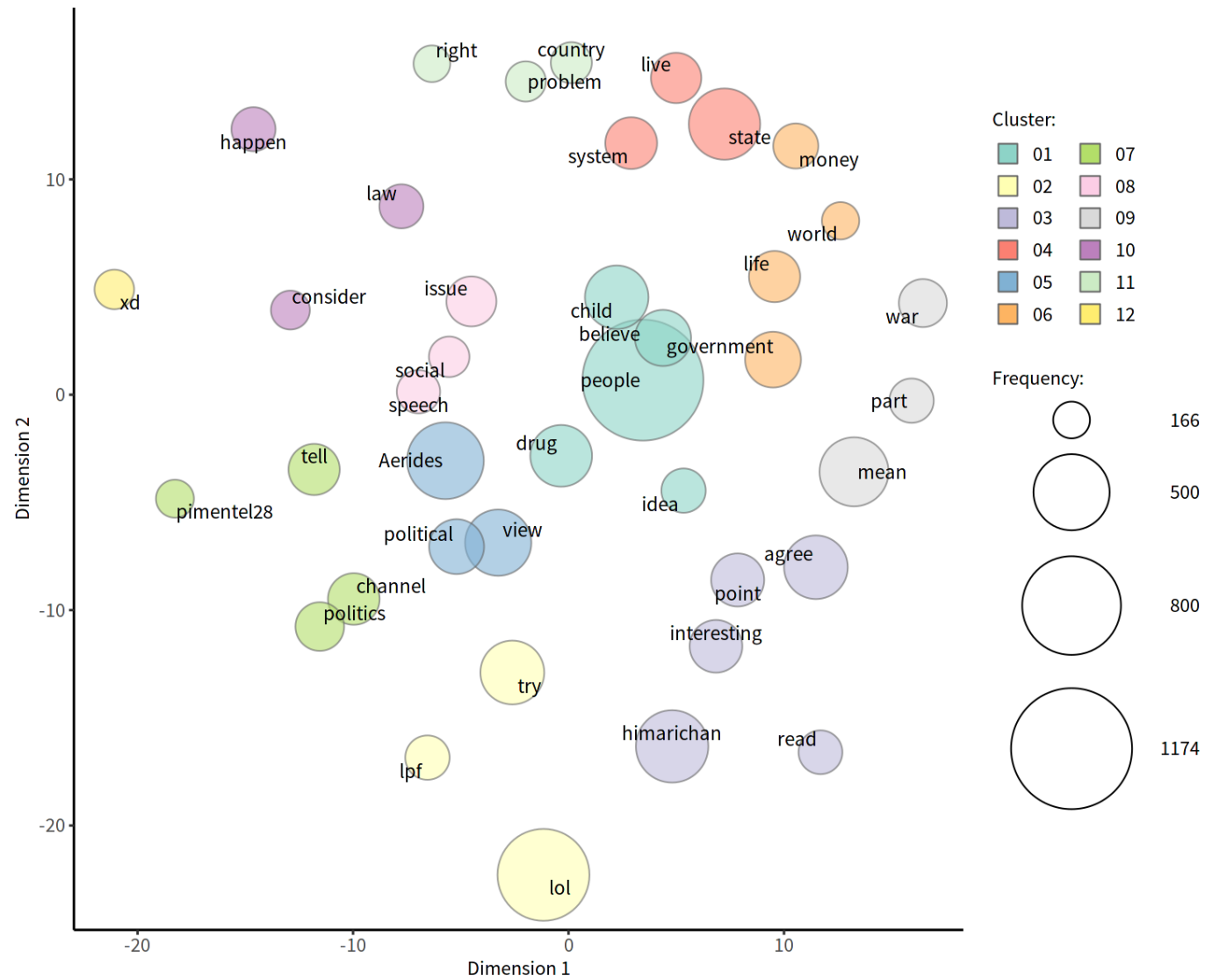
Figure 1a: “Multi-Dimensional Scaling” diagram representing the #libertyinmind IRC channel from October 29th, 2013 to August 4th, 2014 (words = 40, clusters = 11).



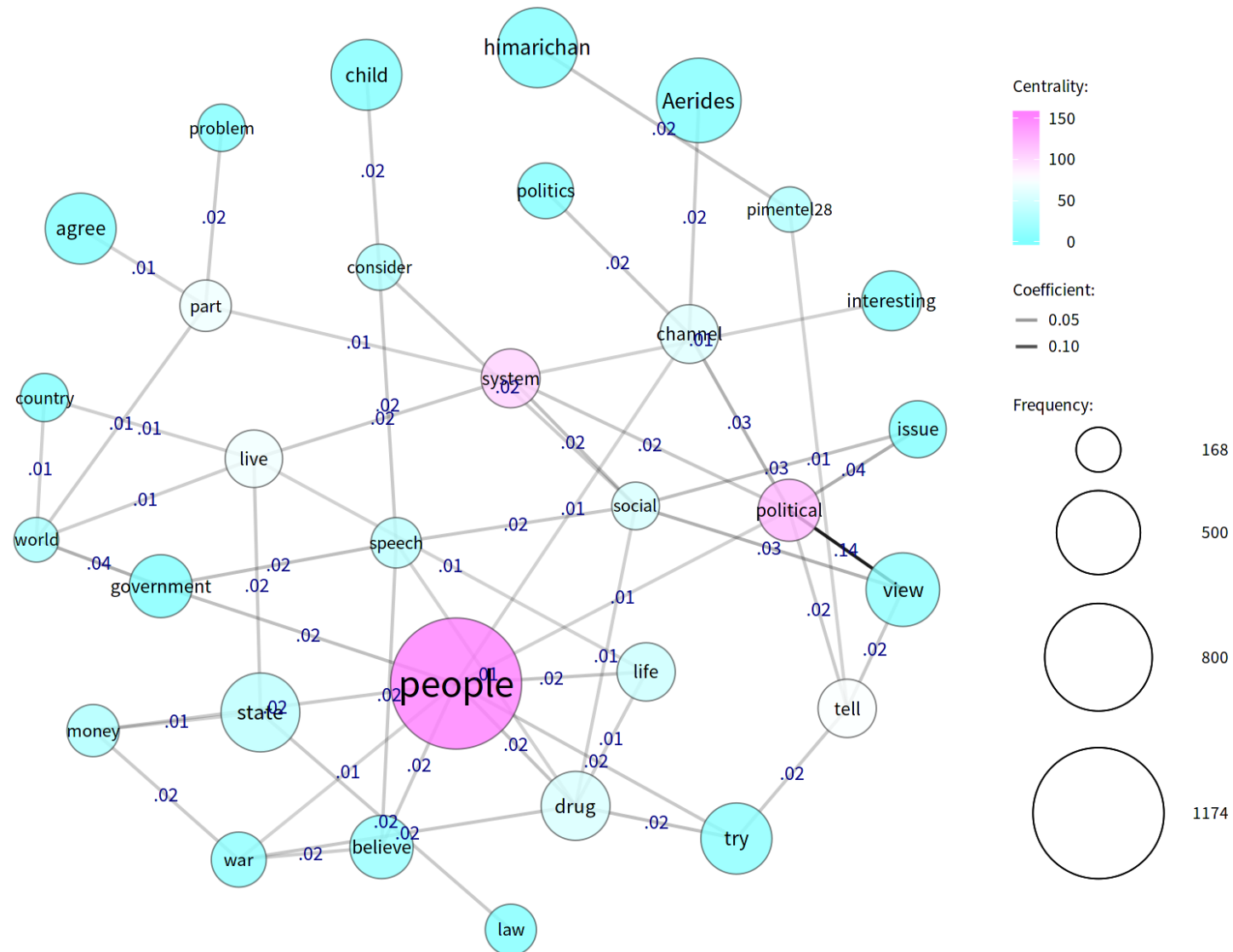
**Figure 1b:** “Co-Occurrence Network of Words” diagram representing the #libertyinmind IRC channel from October 29th, 2013 to August 4th, 2014 (words = 40; edges = 50).



**Figure 2a:** “Multi-Dimensional Scaling” diagram representing the #politics IRC channel from April 29th, 2012 to August 4th, 2014 (words = 40, clusters = 12).

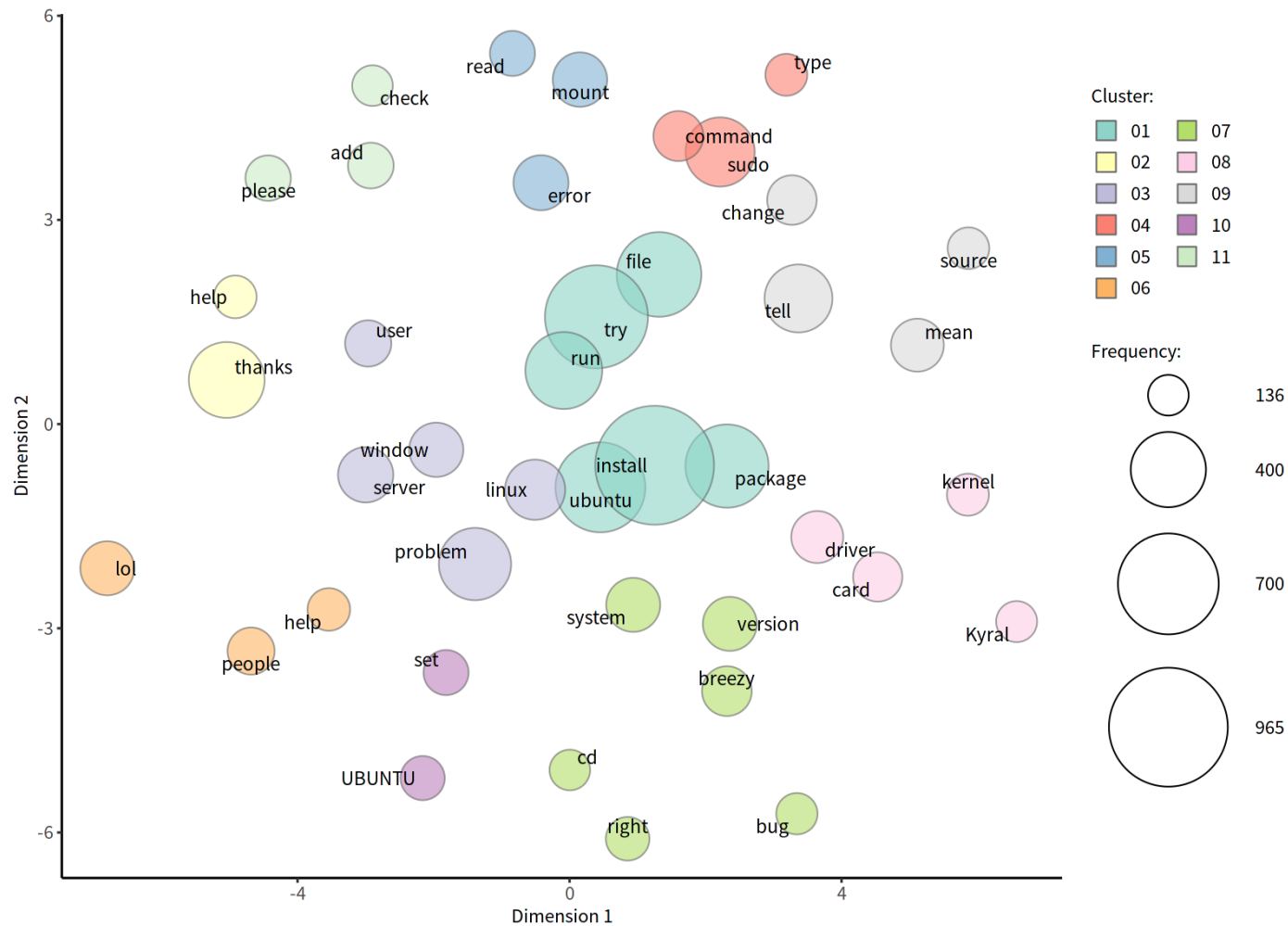


**Figure 2b:** “Co-Occurrence Network of Words” diagram representing the #politics IRC channel from April 29th, 2012 to August 4th, 2014 (words = 40; edges = 50)



## 6. 2 – Diagrammatic Representations of Ubuntu-related channels

**Figure 3a:** “Multi-Dimensional Scaling” diagram representing all IRC channels belonging to the Ubuntu project that were publicly logged on December 25th, 2004 (words = 41, clusters = 11).



**Figure 3b:** “Co-Occurrence Network of Words” diagram representing all IRC channels belonging to the Ubuntu project that were publicly logged on December 25th, 2004 (words = 41; edges = 50).

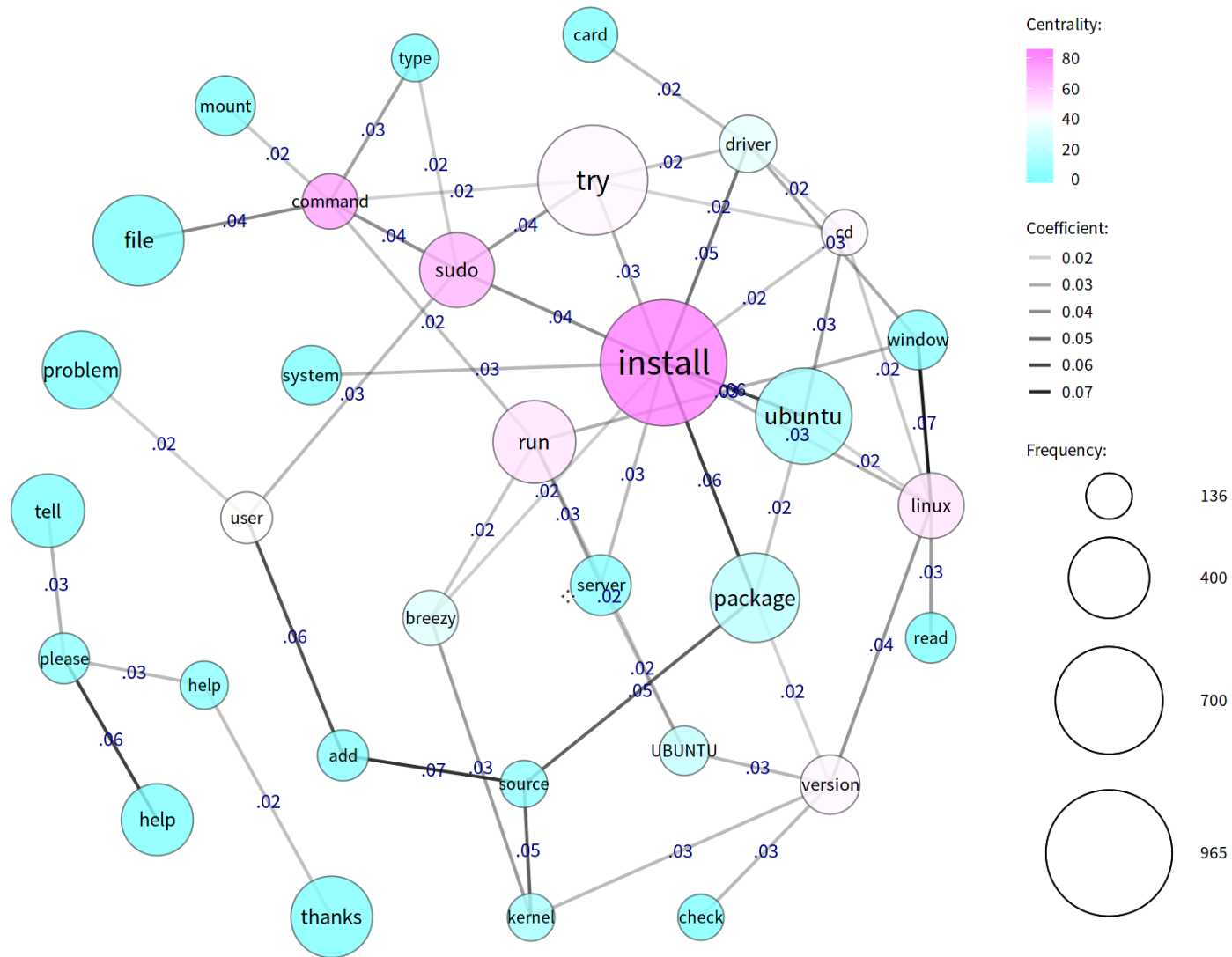
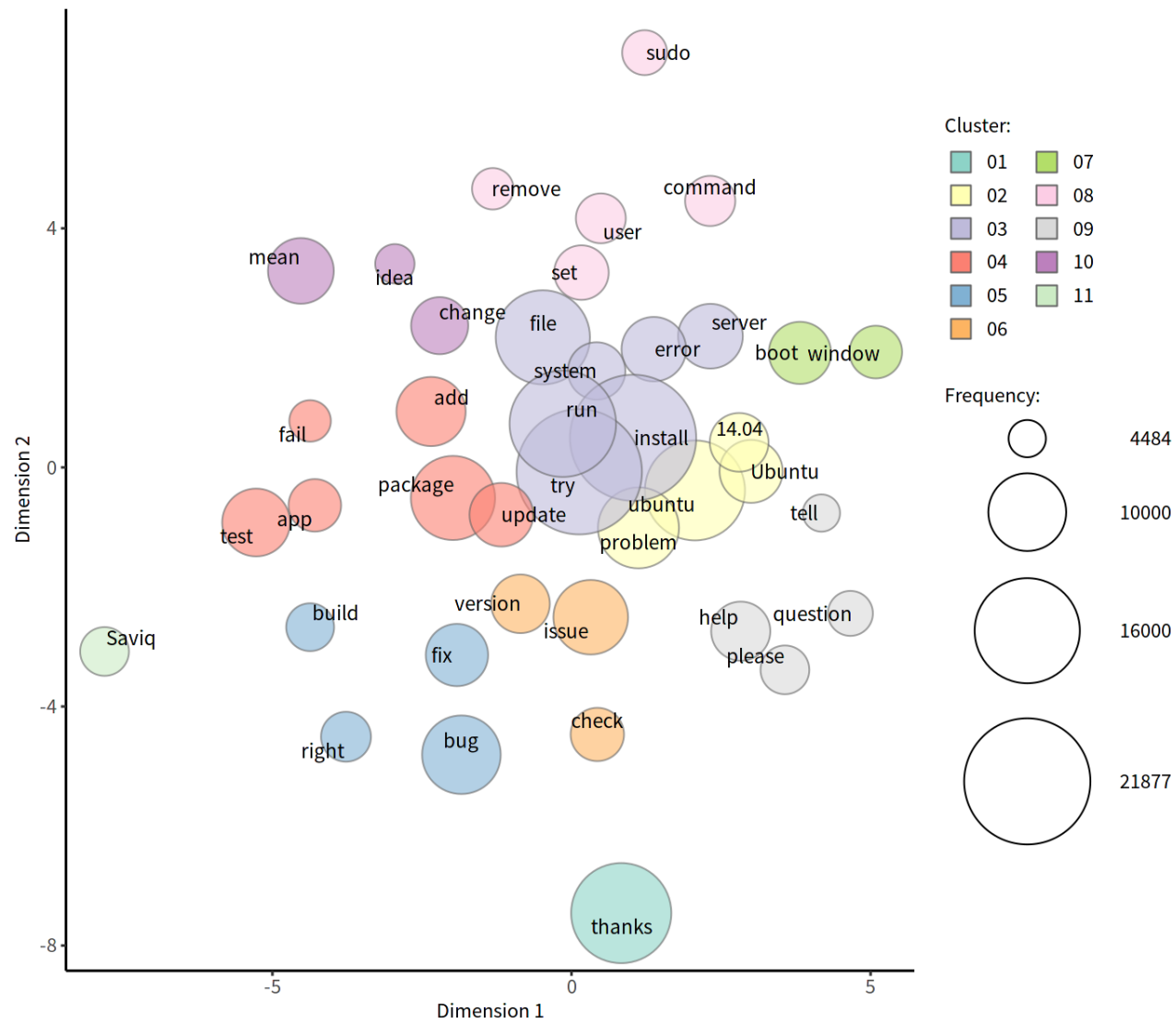


Figure 4a: “Multi-Dimensional Scaling” diagram representing all English-language IRC channels belonging to the Ubuntu project that were publicly logged in July and August of 2014 (words = 40, clusters = 11).



**Figure 4b:** “Co-Occurrence Network of Words” diagram representing all English-language IRC channels belonging to the Ubuntu project that were publicly logged in July and August of 2014 (words = 40; edges = 50).

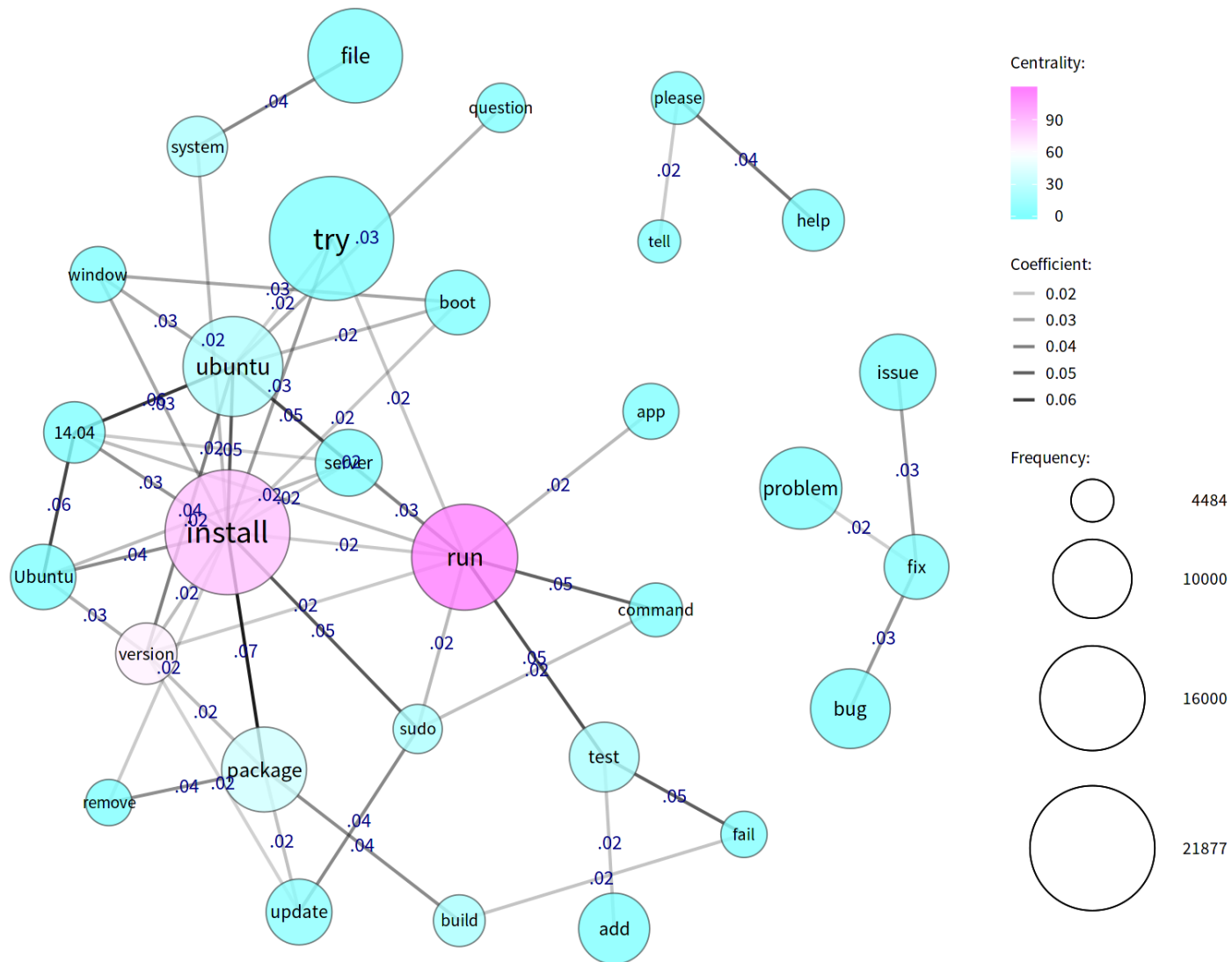


Figure 5a: “Multi-Dimensional Scaling” diagram representing the #kubuntu IRC channel in January of 2013 (words = 41, clusters = 9).

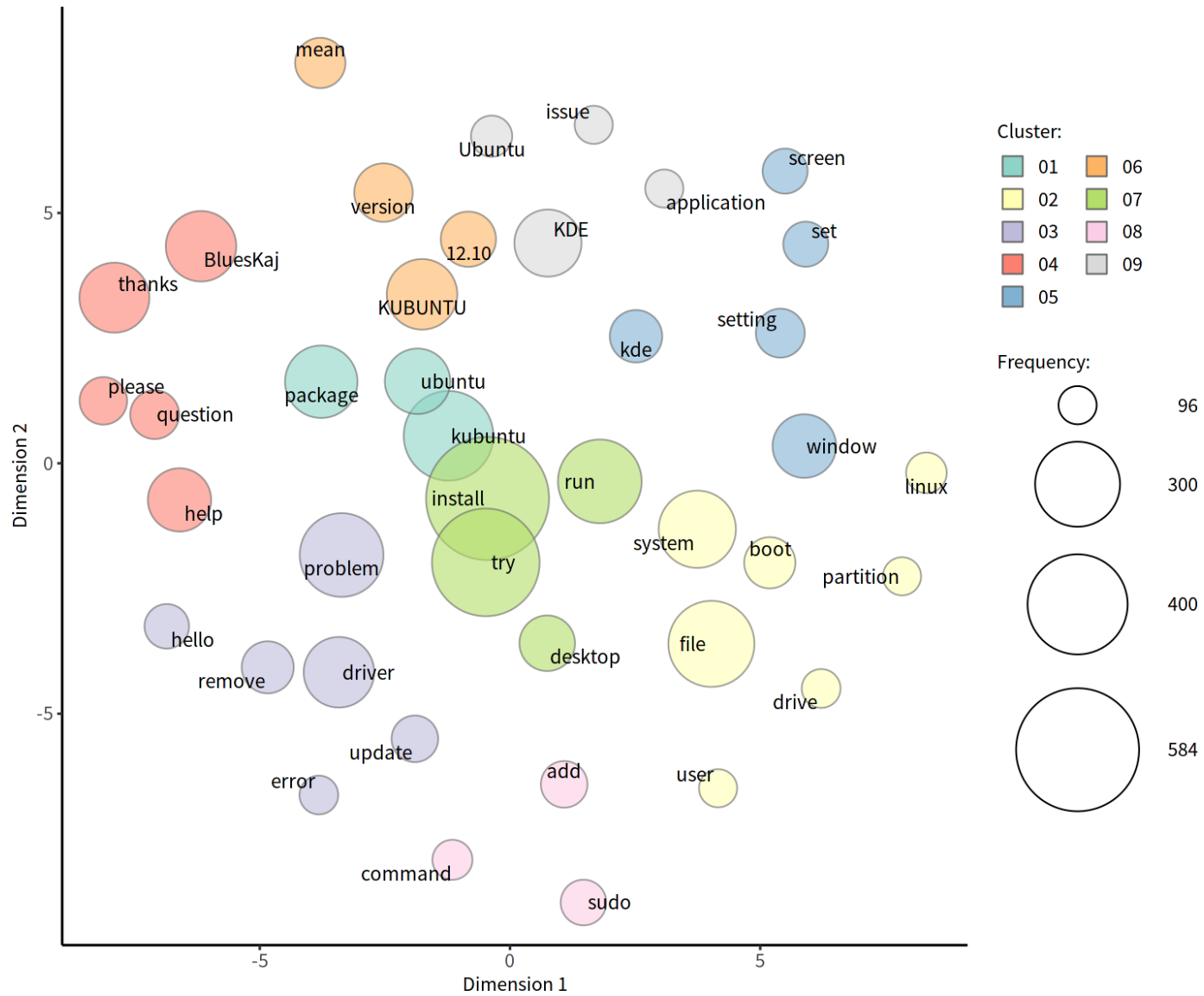
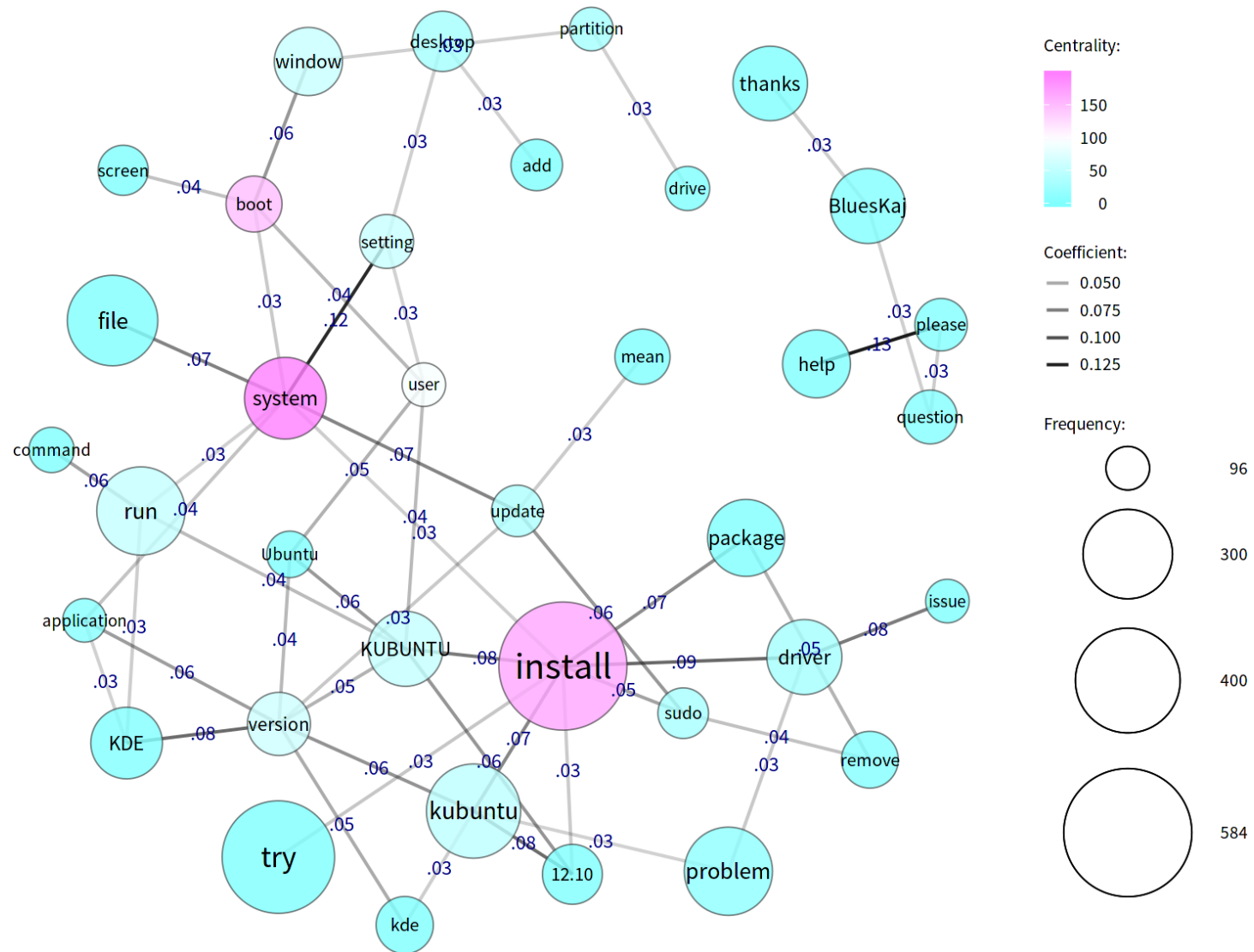
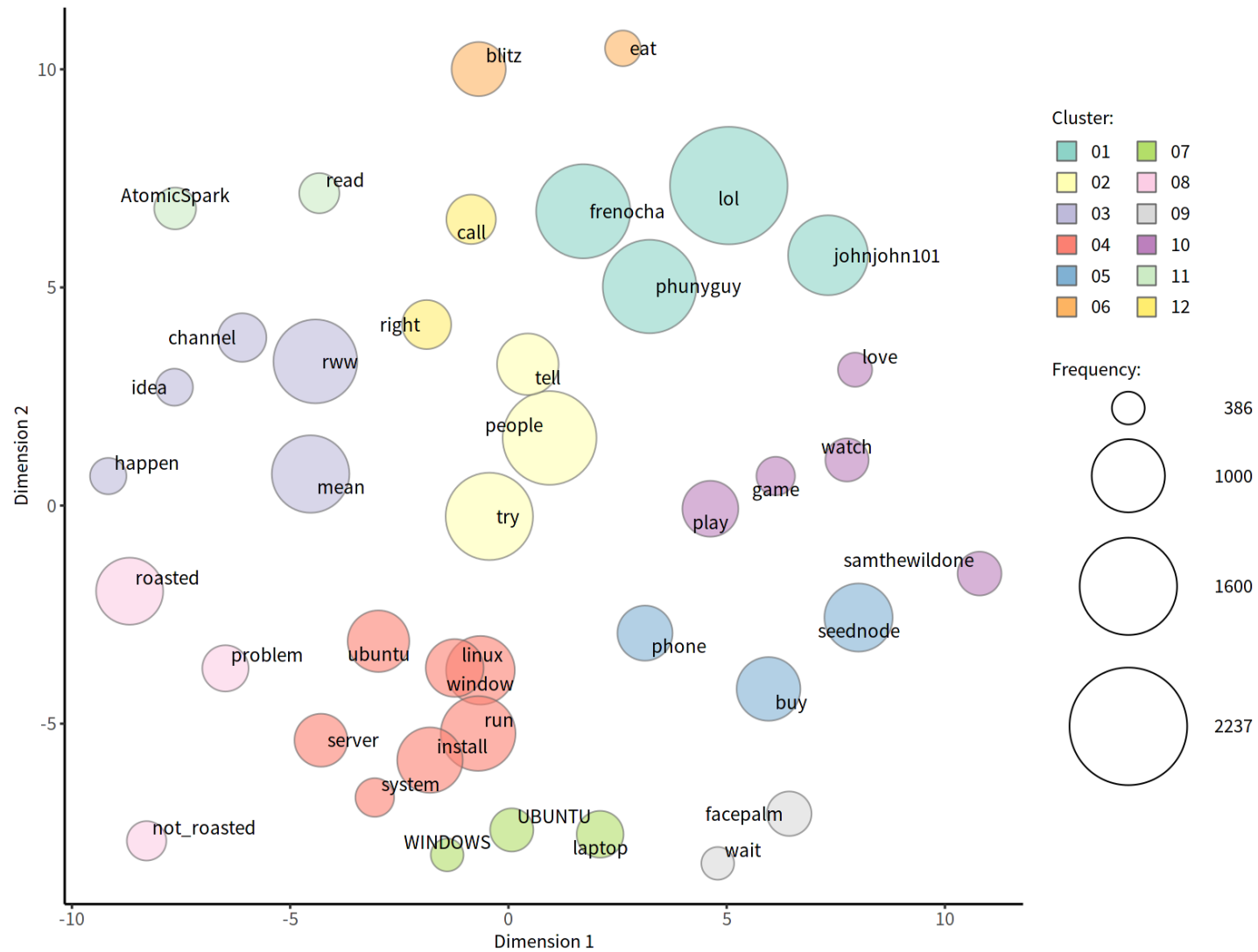


Figure 5b: “Co-Occurrence Network of Words” diagram representing the #kubuntu IRC channel in January of 2013 (words = 41; edges = 50).

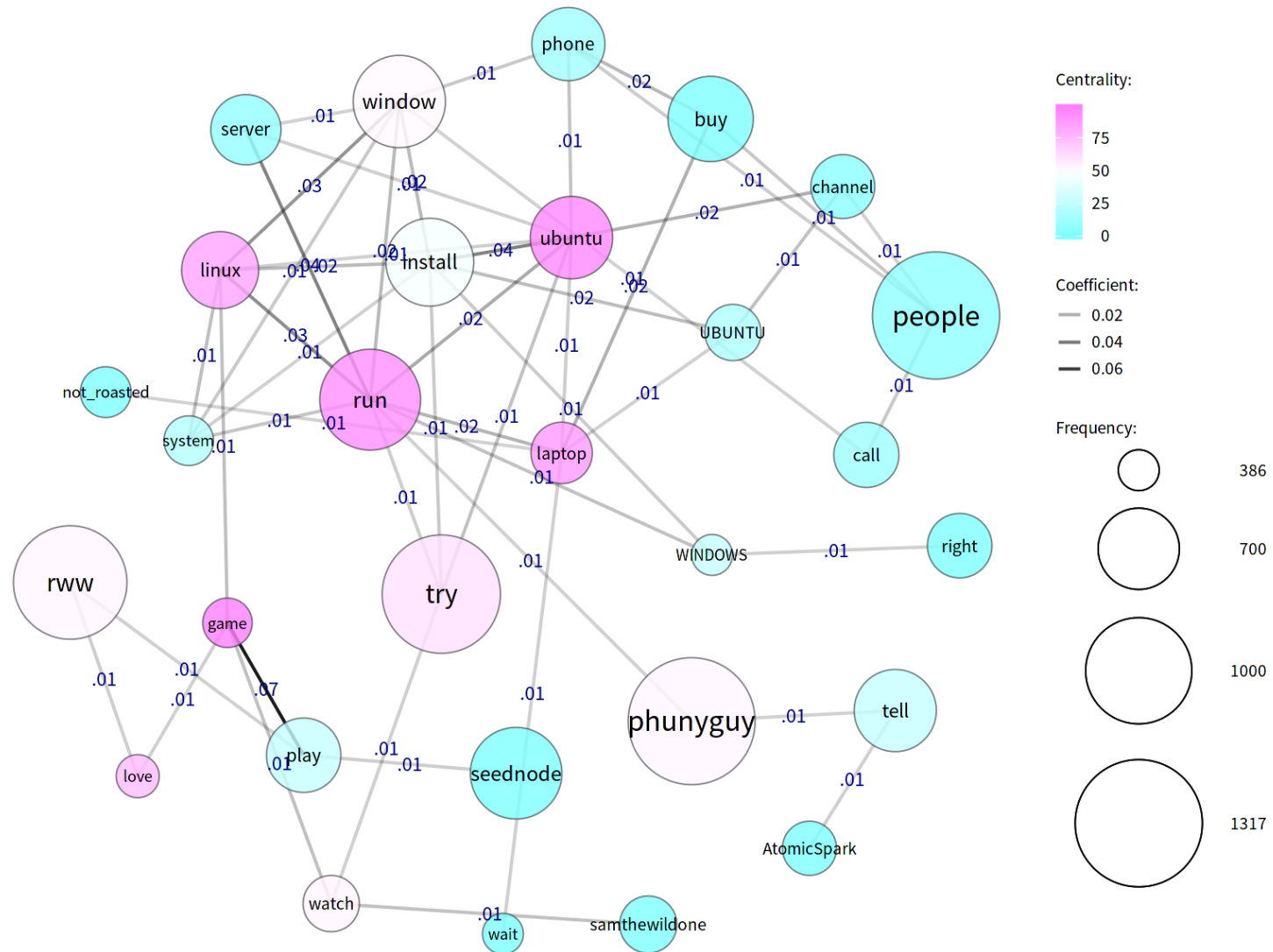


**Figure 6a:** “Multi-Dimensional Scaling” diagram representing the #ubuntu-offtopic IRC channel from August 8th, 2014 to October 13th, 2014

(words = 41, clusters = 12).



**Figure 6b:** “Co-Occurrence Network of Words” diagram representing the #ubuntu-offtopic IRC channel from August 8th, 2014 to October 13th, 2014 (words = 41; edges = 50).



## **7 – References**

*Key references are indicated by asterisks (\*).*

\* Arisbe. *The Peirce Gateway*. – <http://www.iupui.edu/~arisbe/>

Aristotle. *On Interpretation*. ~350 B.C.E. – <http://classics.mit.edu/Aristotle/interpretation.1.1.html>

Aristotle. *Rhetoric - Book 1*. ~350 B.C.E. – <http://classics.mit.edu/Aristotle/rhetoric.1.i.html>

Atkin, A. “Charles Sanders Peirce (1839–1914)”. *Internet Encyclopedia of Philosophy*. – <http://www.iep.utm.edu/peircebi/>

Bush, V. “As We May Think”, *The Atlantic*, July 1945.

Colapietro, V. M. *Peirce's Approach to the Self: A Semiotic Perspective on Human Subjectivity*. State University of New York Press, 1989.

\* Commens. *Digital Companion to C.S. Peirce*. – <http://www.commens.org/>

Cresswell, T. *Place, A Short Introduction*. Blackwell Publishing, 2004.

Deleuze G. *Logique du Sens*, Les Éditions de Minuit, 1969.

Deleuze, G. & F. Guattari. *Mille Plateaux*, Les Éditions de Minuit, 1980.

Derks, D. & A. H. Fischer & A. E. R. Bos. “The role of emotion in computer-mediated communication: A review”. *Computers in Human Behavior*, vol. 24 (3), 2008, 766—785.

Dewes, C. & A. Wichmann & A. Feldmann. “An analysis of Internet chat systems”. *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement*, 2003, 51—64.

\* Doell, W. *Creation and Analysis of a Corpus of Internet Relay Chat*. Thesis deposited as part of the

requirements for the degree of Master of Arts (M.A.) in Applied Language Studies, Carleton University, 2000.

Ducheneaut, N. "Socialization in an Open Source Software Community: A Socio-Technical Analysis". *Computer Supported Cooperative Work (CSCW)*, vol. 14 (4), 2005, 323—368.

Eco, U. *Sémiotique et philosophie du langage*, Presses Universitaires de France, 1988.

Elias, N. *Engagement et distanciation. Contributions à la sociologie de la connaissance*. Fayard, Paris, 1983 (1993).

Esfeld, M. "Le pragmatisme en sémantique et en épistémologie contemporaines". *Philosophia Scientiæ*, 2005, vol. 9 (1), 31—47.

Esquenazi, J.-P. "Éléments pour une sémiotique pragmatique : la situation, comme lieu du sens". *Langage et société*, vol. 80 (1), 1997, 5—38.

Gibson, J. J. "Chapter 8 – The Theory of Affordances". *The Ecological Approach to Visual Perception*, Psychology Press, 1986.

\* Goffman, E. *The Presentation of Self in Everyday Life*. Random House: Anchor Books, 1959.

Goffman, E. *Interaction Ritual: Essays on Face-to-Face Behaviour*. Random House: Pantheon Books, 1967.

Goffman, E. "The Interaction Order: American Sociological Association, 1982 Presidential Address". *American Sociological Review*, 1983, vol. 48(1), 1—17.

Gustaffson, M. *Wittgenstein on using language and playing chess: the breakdown of an analogy, and its consequences*. Paper presented for the Wittgenstein Workshop, 2014.

Haack, S. "We Pragmatists ...: Peirce and Rorty in Conversation". *Manifesto of a Passionate Moderate*, University of Chicago Press, 1998, 31—47.

- Hancock, J. T. & J. Thom-Santelli & T. Ritchie. “Deception and design: The impact of communication technologies on lying behavior”. *Proceedings, Conference on Computer Human Interaction*, vol. 6 (1), 2004, 130—136.
- Hancock, J. T. “Digital deception: Why, when and how people lie online”. *The Oxford Handbook of Internet Psychology*, 2009, 287—301.
- Hancock, J. T. & P. J. Dunham. “Impression Formation in Computer-Mediated Communication Revisited: An Analysis of the Breadth and Intensity of Impressions”. *Communication Research*, vol. 28 (3), 2001, 325—347.
- Heine, S. J. “Chapter 10 – Living In Multicultural Worlds”. *Cultural Psychology (Second Edition)*, W. W. Norton & Company, 2008 (2012), 383—422.
- Heiskala, R. “From Goffman to semiotic sociology”. *Semiotica*, vol. 124 (3–4), 1999, 211—234.
- \* Higuchi, K. *KH Coder (SourceForge)*. – <https://sourceforge.net/projects/khc/>
- Holmer, T. “Discourse Structure Analysis of Chat Communication”. *Language@Internet*, 2008, vol. 5, article 10. – <https://www.languageatinternet.org/articles/2008/1633>
- Kramer, A. D. I. & S. R. Fussell & L. D. Setlock. “Text Analysis as a Tool for Analyzing Conversation in Online Support Groups”. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 2004, 1485—1488.
- Krippendorff, K. H. “Chapter 2 – Conceptual Foundation”. *Content Analysis: An Introduction to Its Methodology (Second Edition)*, Sage Publications, 2004, 18—43.
- Kuhn, T. S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962 (2012).
- Laing, R. D. *Self and Others*. Pelican Books, 1961 (1972).
- Lakatos, I. “Falsification and the Methodology of Scientific Research Programmes”, *Criticism and*

*the Growth of Knowledge*, Cambridge University Press, 1970, 91—196.

Latzko-Toth, G. *La co-construction d'un dispositif sociotechnique de communication : le cas de l'Internet Relay Chat*. Thesis deposited as part of the requirements for the doctorate in communication, Université du Québec à Montréal, 2010.

Lawson, D. *Negotiating Social and Moral Order in Internet Relay Chat*. Thesis deposited as part of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Education, Queensland University of Technology, 2008.

Lévy, P. *La sphère sémantique – Tome 1 : Computation, cognition, économie de l'information*, Hermès, 2011.

\* Liszka, J. J. *A General Introduction to the Semeiotic of Charles S. Peirce*. Indiana University Press, 1996.

Lotman, J. “On the semiosphere”. *Sign System Studies*, vol. 33 (1), 2005.

Manovich, L. “Chapter 1 – What Is New Media?”. *The Language of New Media*, The MIT Press, 2001, 18—61.

McGinn, M. *Routledge Philosophy Guidebook to Wittgenstein and the Philosophical Investigations*. Routledge, 1997.

Mead, G. H. *L'esprit, le soi et la société*. Presses Universitaires de France, 1934 (2006).

\* Michels, M. O. *Real Time Text Analysis on Internet Relay Chat Conversations*. Thesis deposited as part of the requirements for the degree of Master of Science (M.S.) in Information Security, Purdue University, 2012.

Miller, H. *The Presentation of Self in Electronic Life: Goffman on the Internet*. Paper presented at the ‘Embodied Knowledge and Virtual Space’ conference, University of London, 1995.

- Mitkov, R. *The Oxford Handbook of Computational Linguistics (Second Edition)*, Oxford University Press, 2003 (2005).
- Nöth, W. *Handbook Of Semiotics*, Bloomington, Indiana University Press, 1995.
- Oyserman, D. & S. W. S. Lee. "Does Culture Influence What and How We Think? Effects of Priming Individualism and Collectivism". *Psychological Bulletin*, 2008, vol. 134 (2), 311—342.
- Pak, J. *Detecting Deception in Computer-Mediated Communication: A Social Structural Perspective*. Thesis deposited as part of the requirements for the degree of Doctor of Philosophy (Ph.D.) in Information Systems, University of Maryland, 2014.
- Paolillo, J. C. & A. Zelenkauskaitė. "Chapter 5 – Real-time chat". *Pragmatics of Computer-Mediated Communication*, Mouton de Gruyter, 2013, 109—133.
- Papacharissi, Z. "The Presentation of Self in Virtual Life: Characteristics of Personal Home Pages". *Journalism & Mass Communication Quarterly (JMCQ)*, vol. 79 (3), 2002, 643—660.
- Pastinelli, M. "Ethnographie d'une délocalisation virtuelle : Le rapport à l'espace des internautes dans les canaux de 'chat'". *Terminal, technologie de l'information, culture et société*, 1999, vol. 79.
- Pastinelli, M. "Habiter le temps réel : Ethnographie des modalités de l'"être ensemble" dans l'espace électronique". *Anthropologie et Sociétés*, vol. 30 (2), 2006, 199—217.
- \* Peirce, C. S. *The Collected Papers of Charles S. Peirce*. Harvard University Press (Electronic Edition), 1994.
- \* Peirce, C. S. (edited by James Hoopes). *Peirce on Signs*. The University of North Carolina Press, 1991.
- Peirce, C. S. (edited by Patricia Ann Turrisi). *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*. State University of New York Press, 1903 (1997).

- Polanyi, M. *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press, 1958 (1962).
- Popper, K. *The Logic of Scientific Discovery*. Routledge, 1934 / English edition 1958 (2002).
- Queiroz, J. & F. Merrell. “On Peirce's Pragmatic Notion of Semiosis—A Contribution for the Design of Meaning Machines”, *Minds & Machines*, vol. 19 (1), 2009, 129—143.
- Ransdell, J. “Some Leading Ideas of Peirce's Semiotic”, *Semiotica*, vol. 19, 1977, 157—178.
- Ransdell, J. *On the Paradigm of Experience Appropriate for Semiotic*, Delivered at a meeting of the Semiotic Society of America, 1980.
- Ransdell, J. *Teleology and the Autonomy of the Semiosis Process*, Delivered at a conference of the International Association for Semiotic Studies, 1989.
- Reid, E. M. *Electropolis: Communication and Community On Internet Relay Chat*. University of Melbourne, 1991.
- Rheingold, H. *The online community: Homesteading on the Electronic Frontier (revised edition)*. The MIT Press, 1993 (2000).
- Riffe, D. & S. Lacy & F. G. Fico. “Chapter 2 – Defining Content Analysis as a Social Science Tool”. *Analyzing Media Messages: Using Quantitative Content Analysis in Research (Second Edition)*, Lawrence Erlbaum Associates, 2005 (2008), 23—39.
- Riva, G. “The Sociocognitive Psychology of Computer-Mediated Communication: The Present and Future of Technology-Based Interactions”. *CyberPsychology & Behavior*, vol. 5 (6), 2002, 581—598.
- Romanini, V. *Minute Semeiotic*. – <http://www.minutesemeiotic.org/>
- De Saussure, F. *Cours de linguistique générale*. Payot, 1916 (1996).

- Short, T. L. *Peirce's Theory of Signs*. Cambridge University Press, 2007.
- Shapiro, M. "Aspects of a Neo-Peircean Linguistics: Language History as Linguistic Theory", *The Peirce Seminar Papers: Essays in Semiotic Analysis*, vol. 5., Berghahn Books, 2002, vi—224.
- Skagestad, P. "Peirce, Virtuality, and Semiotic", *Philosophy and Cognitive Science (Twentieth World Congress of Philosophy)*, 1998.
- Uszkoreit, H. "What Is Computational Linguistics?". *Hans Uszkoreit Personal Homepage*, Department of Computational Linguistics and Phonetics of Saarland University, 1996 (2000). – [http://www.coli.uni-saarland.de/~hansu/what\\_is\\_cl.html](http://www.coli.uni-saarland.de/~hansu/what_is_cl.html)
- Uthus, D. C. & D. W. Aha. "The Ubuntu Chat Corpus for Multiparticipant Chat Analysis", *AAAI Spring Symposium Series*, 2013.
- Warkentin, D. & M. Woodworth & J. T. Hancock & N. Cormier. "Warrants and Deception in Computer Mediated Communication". *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010, 9—12.
- Wittgenstein, L. *Tractatus Logico-Philosophicus*. 1921 (1922). – <http://www.tractatuslogico-philosophicus.com/>
- \* Wittgenstein, L. *Philosophical Investigations*. Wiley-Blackwell, 1958 (2009).
- Zeng, Q. T. & D. Redd & T. Rindflesch & J. Nebeker. "Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents". *AMIA Annual Symposium Proceedings Archive*, 2012, 1050—1059.

## **8 – Appendixes**

### **8.1 – Appendix 1: Acknowledgements**

- My first thesis adviser Professor Fernando Andacht (University of Ottawa, Canada), for encouraging my auto-didactic proclivities and inspiring me greatly.
- Professor Pierre Lévy (University of Ottawa, Canada), for contributing the initial idea and motivation which gave birth to this research endeavour.
- Professor Koichi Higuchi (Ritsumeikan University of Kyoto, Japan), for his generous help in using his fascinating program *KH Coder*.
- My girlfriend Amanda Estéphan (Université du Québec à Montréal, Canada), for her love, invaluable moral support, and intellectually stimulating input.
- “TrueShiftBlue” (New Zealand), for writing helpful Perl scripts for me and sharing technical insight.
- Jason Sackey (London, United Kingdom), Web developer, for countless interesting discussions and being an amazing IRC friend.
- Professor André Tremblay (University of Ottawa, Canada), for his forthright comments pertaining to multiple aspects of my early work.
- Simon Rivard-Girard (Montréal, Canada), for his assistance in customizing the version of *KH Coder* that was ultimately used to generate the diagrams in this thesis.
- Professor Andrés Pablo Salanova (University of Ottawa, Canada), for his insight regarding the linguistic aspects of my work.
- Everyone else from my family, friends and university who have helped and supported me in this project.

In many ways, this work is a tribute to the countless awesome people from all over the world whom I have met and shared a part of my life with on IRC. Perhaps above anything else, I have attempted to communicate the ways in which my personal involvement in cyberspace has made me who I am.

## **8.2 – Appendix 2: List of Publicly Logged IRC Channels Analyzed**

[1] Ubuntu-related IRC channels (active as of October 2014) – all on the *Freenode* network

Known Non-English or empty channels have been indicated by “\*”. (Some have probably been overlooked.)

|                        |                       |
|------------------------|-----------------------|
| #bzt                   | #ubuntu-ngo           |
| #cloud-init            | *#ubuntu-nl-mwanzo    |
| #juju-dev              | *#ubuntu-nl           |
| #juju-gui              | *#ubuntu-no           |
| #juju                  | #ubuntu-nz            |
| #kubuntu-devel         | #ubuntu-on-air        |
| #kubuntu-se            | #ubuntuone            |
| #kubuntu               | #ubuntu-ops           |
| #launchpad-dev         | *#ubuntu-pe           |
| #launchpad-meeting     | *#ubuntu-ph           |
| #launchpad-reviews     | *#ubuntu-pl           |
| #launchpad             | *#ubuntu-pt           |
| #launchpad-yellow      | *#ubuntu-qa           |
| #lubuntu               | #ubuntu-quality       |
| #maas                  | #ubuntu-release       |
| #ubuntu+1              | *#ubuntu-ro           |
| #ubuntu-app-devel      | #ubuntu-rs            |
| #ubuntu-arb            | *#ubuntu-ru           |
| #ubuntu-arm            | #ubuntu-sa            |
| #ubuntu-artwork        | #ubuntu-scientists    |
| *#ubuntu-ar            | *#ubuntu-se-mote      |
| #ubuntu-autopilot      | #ubuntu-server        |
| #ubuntu-au             | *#ubuntu-se           |
| #ubuntu-bd             | *#ubuntu-si           |
| *#ubuntu-br-sc         | #ubuntustudio-devel   |
| *#ubuntu-br            | #ubuntustudio         |
| #ubuntu-bugs           | *#ubuntu-tn-classroom |
| *#ubuntu-cat           | *#ubuntu-tn-meeting   |
| #ubuntu-ca             | *#ubuntu-tn           |
| *#ubuntu-centroamerica | #ubuntu-toolchain     |
| *#ubuntu-charlas       | #ubuntu-touch         |
| *#ubuntu-ch            | #ubuntu-translators   |
| #ubuntu-ci-eng         | *#ubuntu-tr           |
| #ubuntu-classroom      | *#ubuntu-tw           |
| #ubuntu-cl             | #ubuntu               |
| #ubuntu-cm             | #ubuntu-uds-appdev-1  |
| *#ubuntu-cn            | #ubuntu-uds-appdev-2  |
| #ubuntu-community-team | #ubuntu-uds-client-1  |

---

|                     |                           |
|---------------------|---------------------------|
| *#ubuntu-co         | #ubuntu-uds-client-2      |
| *#ubuntu-cym        | #ubuntu-uds-community-1   |
| #ubuntu-cy          | #ubuntu-uds-community-2   |
| *#ubuntu-cz         | #ubuntu-uds-core-1        |
| *#ubuntu-de-l10n    | #ubuntu-uds-core-2        |
| #ubuntu-design      | #ubuntu-uds-design-1      |
| #ubuntu-desktop     | #ubuntu-uds-devops-1      |
| *#ubuntu-de         | #ubuntu-uds-devops-2      |
| #ubuntu-devel       | #ubuntu-uds-foundations-1 |
| #ubuntu-discuss     | #ubuntu-uds-hallway       |
| #ubuntu-dk-moede    | #ubuntu-uds-platform-1    |
| *#ubuntu-dk         | #ubuntu-uds-platform-2    |
| #ubuntu-doc         | #ubuntu-uds-plenary       |
| *#ubuntu-dz         | #ubuntu-uds-servercloud-1 |
| #ubuntu-eg          | #ubuntu-uds-servercloud-2 |
| *#ubuntu-es         | #ubuntu-uds               |
| #ubuntu-et          | #ubuntu-uds-users-1       |
| *#ubuntu-fi         | #ubuntu-uds-users-2       |
| #ubuntu-for-all     | #ubuntu-uk                |
| *#ubuntu-fr-l10n    | #ubuntu-unity             |
| #ubuntu-gnome       | #ubuntu-us-ar             |
| *#ubuntu-gr         | #ubuntu-us-ca             |
| *#ubuntu-hr         | #ubuntu-us-co             |
| #ubuntu-ie          | #ubuntu-us-dc             |
| *#ubuntu-il         | #ubuntu-us-de             |
| #ubuntu-installer   | #ubuntu-us-fl             |
| #ubuntu-irc         | #ubuntu-us-ga             |
| *#ubuntu-ir         | #ubuntu-us-md             |
| *#ubuntu-it-meeting | #ubuntu-us-mi             |
| *#ubuntu-it         | #ubuntu-us-mn             |
| #ubuntu-java        | #ubuntu-us-nc             |
| *#ubuntu-jp         | #ubuntu-us-oh             |
| #ubuntu-kernel      | #ubuntu-us-or             |
| *#ubuntu-ke         | #ubuntu-us-pa             |
| *#ubuntu-ko         | #ubuntu-us-sc             |
| #ubuntukylin-devel  | #ubuntu-us-tn             |
| #ubuntu-l10n-es     | #ubuntu-us-tx             |
| *#ubuntu-lb         | #ubuntu-us-wa             |
| #ubuntu-leadership  | *#ubuntu-uy               |
| #ubuntu-locoteams   | #ubuntu-vn                |
| #ubuntu-meeting-2   | #ubuntu-website           |
| #ubuntu-meeting     | #ubuntu-women-project     |
| #ubuntu-mir         | #ubuntu-x                 |
| #ubuntu-mk          | #ubuntu-youth             |
| #ubuntu-motu        | #ubuntu-za                |
| #ubuntu-mozillateam | #upstart                  |

---

|                |                |
|----------------|----------------|
| #ubuntu-mythtv | #xubuntu-devel |
| #ubuntu-news   | #xubuntu       |

[2] MediaWiki- and Wikimedia-related IRC channels (active as of October 2014) – all on the *Freenode* network

#mediawiki  
#mediawiki-visualeditor  
#semantic-mediawiki  
#wikidata  
#wikimedia-analytics  
#wikimedia-dev  
#wikimedia-fundraising  
#wikimedia-labs  
#wikimedia-mobile  
#wikimedia-multimedia  
#wikimedia-office  
#wikimedia-operations  
#wikimedia-research  
#wikipedia\_usability

### **8.3 – Appendix 3: Data Pre-Processing**

Quantitative analysis of IRC chat-logs presents many challenges for the researcher. The format of complete “logs”<sup>167</sup> is such that much of the information displayed consists of automatically generated “events” of little significance with regard to text analysis such as “joins”, “parts”, “nick” and “topic” changes... which entails that pre-processing the data is no trivial task for the non-programmer. Furthermore, chat-rooms – especially larger ones – tend to be rather chaotic. On most Ubuntu-related channels, people frequently speak many languages and routinely insert “garbage” in the conversation,<sup>168</sup> careful moderation of such large channels is of paramount importance. On channels related to projects such as MediaWiki and Wikimedia, “bots” are so heavily involved in the conversations, and humans sometimes so little by comparison, that their robotic utterances can make up the whole of the “chat” for entire days! In a similar line of thought, lines of programming code shared in programming-related channels do not lend themselves well to natural language processing. For all those reasons among others, before their statistical analysis, I have heavily processed the raw logs to preserve only data judged relevant for the present research.

Several scripts written in the Perl programming language were used to pre-process the data so as to make its analysis feasible<sup>169</sup>. A first script was used to parse “raw” IRC logs in order to prune everything but the nicknames and messages. Due to slight differences in the format of the logs made publicly available by the two projects, two variants of the same script were required: one for logs of

167 The structure of which is illustrated by a concrete example in [8.14 – Appendix 14: IRC Interface Example \(Quassel Client\)](#).

168 Typically *spam*, an exceptionally blatant example of which can be found in [8.16 – Appendix 16: IRC Log Sample – Spam in #ubuntu-offtopic](#).

169 All the scripts used for that purpose can be found in [8.4 – Appendix 4: Various Perl Scripts Used for Pre-Processing](#).

channels related to Ubuntu, and one for logs of channels related to MediaWiki and Wikimedia. A second script was subsequently used to strip the data of all non-ASCII characters for the purpose of facilitating *Part-Of-Speech Tagging*. For the same purpose, in some cases, a third script was applied to delete all lines containing over 50% of non-alphabetic characters. For some parts of my analysis, a fourth script was used to delete the nicknames of the speakers, preserving only their messages. Finally, in the case of the exhaustive analysis of data from MediaWiki, Wikimedia and related projects, I used a simple Perl one-liner to delete all lines by known “bots”, as this data cluttered the visualizations.

To avoid pre-processing errors due to limitations inherent to KH Coder, <, > and \ characters remaining in the log files were removed automatically with the “Replace All” function of a simple text editor. Finally, in some cases<sup>170</sup>, a handful of lines had to be removed manually because the Stanford Part-of-Speech Tagger could not process them (some words in foreign languages, exotic characters, garbage text, non-printing characters, etc.). Methodologically, they may be written off as statistically irrelevant due to the sheer volume of data processed. The same can be said of the few parts of my own logs that I never recorded due to power outages or computer crashes<sup>171</sup>.

170 For analyzes performed after the end of September 2014, I managed to find a work-around for this issue with the kind help of Prof. Koichi Higuchi. I did not judge necessary to repeat previous analyzes to include the deleted lines.

171 This is relevant to my analysis of the #libertyinmind, #politics, and #ubuntu-offtopic data-sets.

## **8.4 – Appendix 4: Various Perl Scripts Used for Pre-Processing**

[1a], [1b] and [4] courtesy of TrueShiftBlue. [2] adapted from a line found on an Internet forum. [3] courtesy of Prof. Koichi Higuchi. [5] adapted from a line found on the StackOverflow website.

[1a] *Parsing “raw” public logs of channels related to Ubuntu in order to prune everything but the nicknames and messages:*

```
perl -nwE 'print "$1 $2\n" if /^\[0-9]{2}:\[0-9]{2}\] <([>]+)>
(.*)$/;' /path/to/logs.txt > /path/to/output.txt
```

[1b] *Variant of [1a] for MediaWiki- and Wikimedia-related channels:*

```
perl -nwE 'print "$1 $2\n" if /^\[0-9]{2}:\[0-9]{2}:\[0-9]{2}\]
<([>]+)> (.*)$/;' /path/to/logs.txt > /path/to/output.txt
```

[2] *Deletion of all non-ASCII characters:*

```
perl -i.bk -pe 's/[^\[:ascii:]]//g;' file.txt
```

[3] *For larger, more complex logs: deletion of lines containing over 50% of non-alphabetic characters. The script was saved as 'prune-non-alphabetic.pl'; usage is:*

```
perl prune-non-alphabetic.pl /path/to/logs.txt > /path/to/output.txt
```

*The script itself:*

```
#!/usr/bin/perl -w

use strict;
use warnings;

my $file = $ARGV[0];

unless (-e $file){
    print "no such file: $file\n";
    exit;
}

open my $fh, '<', $file or die;
while (<$fh>){
    chomp;
    my $length = length($_);
    next unless $length;
```

```
my @alpha = ($_ =~ /[A-Z]/iog);  
if ( ($#alpha + 1) / length >= 0.5 ){  
    print "$_\n";  
}  
}
```

[4] *Deletion of the first word of each line so as to delete the nicknames of participants, keeping only the messages.*

```
perl -nwE 's/\S+\s//; print "$_";' /path/to/logs.txt >  
/path/to/output.txt
```

[5] *Deletion of all lines that begin by WORD, primarily to delete all lines by known bots.*

```
perl -ni.bk -E "print unless /^WORD/" /path/to/logs.txt
```

**8.5 – Appendix 5: List of Stopwords Used**

|          |           |             |
|----------|-----------|-------------|
| 0        | around    | differently |
| 1        | as        | do          |
| 2        | ask       | dont        |
| 3        | at        | down        |
| 4        | away      | during      |
| 5        | b         | e           |
| 6        | back      | each        |
| 7        | bad       | either      |
| 8        | be        | end         |
| 9        | because   | enough      |
| a        | before    | even        |
| about    | begin     | evenly      |
| above    | behind    | ever        |
| across   | being     | every       |
| actually | below     | everybody   |
| after    | best      | everyone    |
| again    | better    | everything  |
| against  | between   | everywhere  |
| ah       | big       | f           |
| all      | bit       | far         |
| almost   | both      | feel        |
| alone    | but       | few         |
| along    | by        | find        |
| already  | c         | first       |
| also     | can       | for         |
| although | case      | four        |
| always   | certain   | from        |
| among    | certainly | full        |
| an       | clear     | fully       |
| and      | clearly   | further     |
| another  | come      | g           |
| any      | could     | get         |
| anybody  | course    | girl        |
| anyone   | d         | give        |
| anything | day       | go          |
| anyway   | differ    | good        |
| anywhere | different | great       |
| guess    | latest    | nothing     |

---

|         |        |           |
|---------|--------|-----------|
| guy     | least  | now       |
| h       | less   | nowhere   |
| have    | let    | o         |
| he      | lets   | of        |
| hear    | like   | off       |
| her     | likely | often     |
| here    | long   | oh        |
| hers    | look   | ok        |
| herself | lot    | okay      |
| hi      | m      | old       |
| high    | make   | on        |
| him     | man    | once      |
| himself | many   | one       |
| his     | may    | only      |
| hour    | maybe  | or        |
| how     | me     | other     |
| however | men    | ought     |
| http    | might  | our       |
| I       | more   | ourselves |
| i       | most   | out       |
| if      | mostly | over      |
| im      | mr     | own       |
| in      | mrs    | p         |
| indeed  | much   | per       |
| into    | must   | perhaps   |
| it      | my     | person    |
| its     | myself | place     |
| itself  | n      | present   |
| j       | need   | pretty    |
| just    | never  | probably  |
| k       | new    | put       |
| keep    | next   | q         |
| kind    | nice   | quite     |
| know    | no     | r         |
| l       | nobody | rather    |
| large   | none   | really    |
| largely | noone  | s         |
| last    | nor    | same      |
| later   | not    | say       |
| second  | this   | with      |
| see     | those  | within    |

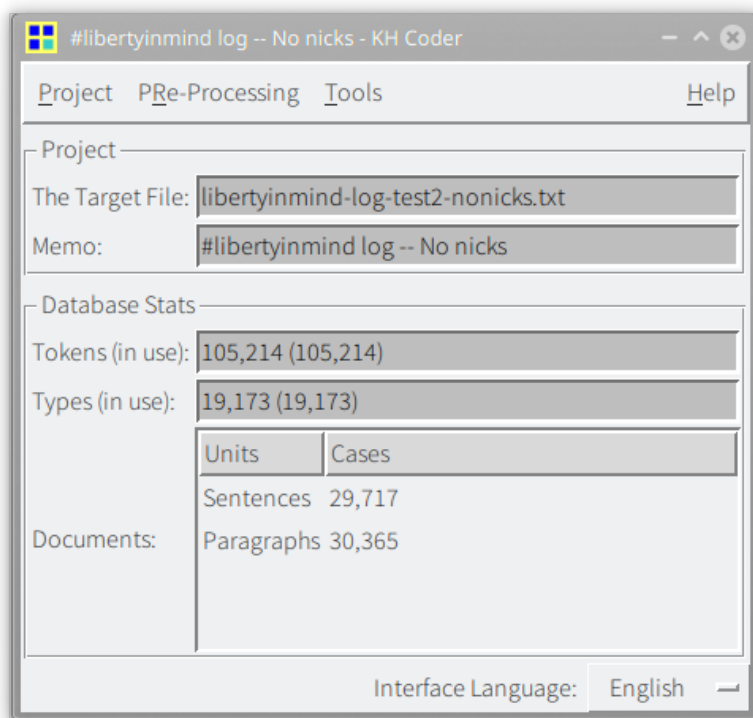
---

|            |          |          |
|------------|----------|----------|
| seem       | though   | without  |
| several    | three    | work     |
| shall      | through  | worse    |
| she        | thus     | worst    |
| shit       | time     | would    |
| should     | to       | x        |
| show       | today    | y        |
| side       | together | yeah     |
| since      | too      | year     |
| small      | toward   | yep      |
| so         | towards  | yes      |
| some       | two      | yet      |
| somebody   | u        | you      |
| someone    | under    | your     |
| something  | until    | youre    |
| sometimes  | up       | yourself |
| somewhere  | upon     | z        |
| start      | us       |          |
| still      | use      |          |
| stuff      | v        |          |
| such       | very     |          |
| sure       | w        |          |
| t          | want     |          |
| take       | way      |          |
| talk       | we       |          |
| than       | well     |          |
| that       | what     |          |
| the        | when     |          |
| their      | where    |          |
| them       | whether  |          |
| themselves | which    |          |
| then       | while    |          |
| there      | who      |          |
| therefore  | whole    |          |
| these      | whom     |          |
| they       | whose    |          |
| thing      | why      |          |
| think      | will     |          |

## **8.6 – Appendix 6: KH Coder Interface Example**

To facilitate comparisons, the parameters used were identical for all analyzes performed, with the notable exceptions of the minimum word count for word selection, which was chosen according to the size of each log in order to keep the number of words visualized to something stable and intuitively understandable ( $40 \pm 1$ ), and the number of clusters in “Multi-Dimensional Scaling” diagrams, which was chosen according to which groupings appeared to make the most sense. Otherwise, the number of edges was different in the case of a single “Co-Occurrence Network of Words”, [Figure 8d](#).

*The image below is an example of the main window of KH Coder’s user interface, as it appears with the #libertyinmind text data open (after pre-processing):*



The following settings were used for the “Multi-Dimensional Scaling” diagram representing *#libertyinmind* (*Figure 1a*):

Options: Multi-Dimensional Scaling of Words

Select the unit & words

Unit: Paragraphs

Filter words by Term Frequency

Min. TF: 148 Max. TF:

Filter words by Document Frequency

Min. DF: 1 Max. DF:

Filter words by POS

- ☒ Noun
- ☒ ProperNoun
- ☒ Foreign
- ☒ PRP
- ☒ Adj
- ☒ Adv
- ☒ Verb
- ☒ W
- ☒ OTHER

All Default Clear

Number of selected words

Check 40

Multi-Dimensional Scaling Options

Method: Kruskal Distance: Jaccard

Dimensions: 2 (from 1 to 3)

☒ Bubble plot: Size of bubbles 120 %

☒ Standardize the size: Variance 155 %

☒ Indicate clusters by different colors

Number of clusters: 11 (from 2 to 12)

☒ Adjacent clusters

☒ Translucent Colors (not suitable for EMF/EPS)

☒ Fix aspect ratio

☒ Use random starts to find better solutions  
\*It may take 10 minutes or more.

Font size: 115 % ☐ Bold Plot size: 1200

☐ Keep this window open

OK Cancel

The following settings were used for the “Co-Occurrence Network of Words” diagram representing #libertyinmind (*Figure 1b*):

**Options: Co-Occurrence Network of Words**

**Select the unit & words**

Unit: Paragraphs

Filter words by Term Frequency  
Min. TF: 148 Max. TF:

Filter words by Document Frequency  
Min. DF: 1 Max. DF:

Filter words by POS

- ☒ Noun
- ☒ ProperNoun
- ☒ Foreign
- ☒ PRP
- ☒ Adj
- ☒ Adv
- ☒ Verb
- ☒ W
- ☒ OTHER

All Default Clear

Number of selected words  
Check 40

**Co-Occurrence Network Options**

Type of edges:  
☒ words - words 
 ☐ words - variables / headings

Variable / Headings: n / a

Filter edges: Cosine

☒ Top 50 
 ☐ Coef. >= 0.2

☒ Darker lines for stronger edges 
 ☒ Show coef.

☒ Bubble plot: Size of bubbles 125 %

☒ Standardize the size: Variance 140 %

☒ Larger nodes for higher frequency words

☒ Variable font size

☐ Smaller Nodes

☐ Highlight the minimum spanning tree

☐ Draw the minimum spanning tree only

☐ Color: correlation with variables / position

Position

☒ Avoid overlapping of labels

☒ Translucent Colors (not suitable for EMF/EPS)

☐ Grayscale (centralities & communities)

Font size: 100 % ☐ Bold Plot size: 1200

☐ Keep this window open

OK Cancel

### **8.7 – Appendix 7: Ludwig Wittgenstein: *Philosophical Investigations***

The philosophy of Peirce and Wittgenstein prominently involves pondering *signification* or *meaning*: in Peirce's case mainly through his theory of *signs*, in Wittgenstein's case mainly through his study of *language*<sup>172</sup>, but both by assuming a fundamental continuity between their subject matter and the rest of the world – or as Wittgenstein wrote: “Language (or thought?) is something unique’ – this proves to be a superstition (not a mistake!), itself produced by grammatical illusions.” [PI §110] It occurs to me that together, the two influential thinkers provide a strong foundation for studying the nature of mind, language and communication.

Although I wish to avoid the most intricate aspects of exegesis, one major interpretative conundrum that is relevant to the present inquiry is the extent to which the thought of the later Wittgenstein (*Philosophical Investigations*, 1953<sup>173</sup>) differs from that of the early Wittgenstein (*Tractatus Logico-Philosophicus*, 1921<sup>174</sup>) – in other words, contemporary scholarship disagrees as to whether the philosopher's thought ought to be viewed as “continuous” or “discontinuous”<sup>175</sup>. The arguments I will discuss and put to use in this work are for the most part those of the later Wittgenstein, who explicitly repudiated many core tenets of his earlier philosophy; in many regards, the *Investigations* consist in a critique of the *Tractatus*. However, my analysis will concurrently build upon elements of continuity and discontinuity in the philosopher's thought.

172 From a Peircean perspective, Wittgenstein's account of meaning applies specifically to the *symbols* of human language (understood very generally). His definition of a “symbol”, introduced in the *Tractatus* (and apparently retained in the *Investigations*, although little use is made of the concept), is in principle compatible with that of Peirce: “to recognize the symbol in the sign we must *consider the significant use*” [TLP 3.326] (emphasis mine). On the other hand, the word “sign”, defined in the *Tractatus* as “what can be perceived of a symbol” [TLP 3.32], corresponds roughly to what Peirce would call the instantiation of a symbol as a *sinsign* or *token*.

173 Hereafter abridged as *Investigations*.

174 Hereafter abridged as *Tractatus*.

175 This situation appears similar to that found in diverging contemporary interpretations of Peirce's thought.

In the *Investigations*, Wittgenstein's inquiry into human language (and its meaning) prompted him to suggest, and reject, many conceivable pictures of its essence, starting – in the very first paragraph of the book – with that of Augustine, which is substantially similar to that put forward by Wittgenstein himself in the *Tractatus*. That conception of the way language relates to the world, which may be termed the “name—object” model [Johnson, 2008, p. 2], may be summarized as the idea that “the words in language name objects – sentences are combinations of such names” [PI §1]<sup>176</sup>. That is to say, the components of language correspond to individual things (“simples”), while propositions express the compounding of those things (“composites”).

In the *Tractatus*<sup>177</sup>, Wittgenstein concurred with Augustine in saying that the “name means the object” [TLP 3.203]<sup>178</sup>. However, Wittgenstein also argued that “only in the context of a proposition has a name meaning” [TLP 3.3]; the ability of individual words to refer to *states of affairs*<sup>179</sup> or *facts* of reality depends upon their “relations to one another” [TLP 4.4661] sharing the logical structure of what they represent. In turn, “[a] proposition is the expression of its truth-conditions” [TLP 4.431], namely *what must be the case for the proposition to be true*. In other words, two propositions are logically equivalent if and only if they are true (and, incidentally, false) in the same range of possible

176 We find here a *nominalistic* conception, which in Peircean terms denies the existence of “names” (as defined above) *qua* symbols, in that sense having as philosophical antithesis “realism”, although that latter concept has other uses in philosophy (notably opposing it to “anti-realism”, or to “idealism” and “solipsism” – see [footnote #44](#) in the section [2. 1 \(c\) – Semiotic & pragmatism as interdependent frames of reference](#) and [footnote #181](#) later in this same section).

177 A work which attempted to set out the conditions for a *logically perfect (or ideal) language*, described by Bertrand Russell as one in which “there [would] be one word and no more for every simple object, and everything that is not simple [would] be expressed by a combination of words” [1956/1972, as quoted in the *Stanford Encyclopedia of Philosophy*].

178 Or, as Russell put it in the book's introduction: “The possibility of a proposition representing a fact rests upon the fact that in it objects are represented by signs.”

179 See also: “The simplest kind of proposition, an elementary proposition, asserts the existence of a state of affairs.” [TLP 4.21]

states of affairs. According to the early Wittgenstein, knowing the truth-conditions of a proposition is knowing its meaning, and to explain this meaning, one must *analyze* the proposition by breaking it down into its constituent “simples” until the “one and only [...] complete analysis of the proposition” [TLP 3.25] is obtained, revealing the proposition's logical structure.

In the *Investigations*, Wittgenstein severely criticizes his earlier conception of language. Thus, he calls attention to the fact that “[we] use the word 'composite' (and therefore the word 'simple') in an enormous number of different and differently related ways.” [PI §47] It is commonly unclear whether a signifying unit is one or the other. Furthermore, the very fact that “complex” symbols are *not* decomposed into “simple” parts often seems fundamental to their meaning: nothing indicates that telling someone to bring “the broomstick and the brush which is fitted on to it” is equivalent to, let alone *superior* to, telling that person to bring “the broom” [PI §60].

In addition, any interpretation of the 'name—object' model calls for an account of the process of *naming*, by way of which a name becomes connected to an object – without which it is unintelligible. Wittgenstein first provides the following tentative attempt at elucidating that process:

The word 'signify' is perhaps most straightforwardly applied when the name is actually a mark on the object signified. [...] When philosophizing, it will often prove useful to say to ourselves: naming something is rather like attaching a name tag to it. [PI §15]

This analogy rapidly shows unacceptable limitations<sup>180</sup> when we attempt to apply it to “things” that are not *objects*. Although we do attach actual name tags to well-defined “spaces” that are not objects in a strict sense (such as brands, countries, or even online communities), where are we to put the

180 Acknowledged by Wittgenstein, who evidently never intended the attempt in §15 to be anything more than “*often* [...] useful” (emphasis mine).

name tags of the purple, the good and the beautiful? On every *example* of such-and-such that we can find? – But how are we to choose the examples *in the first place*?

Thus, the previous account, although it may be useful to describe the process of naming in the case of the least ambiguously observable *particular* terms (most evidently proper nouns), is of relatively little use to describe how *general* terms come to signify<sup>181</sup> (this applies to the greatest part of ordinary uses of language-signs).

Making the “name tag” a transient utterance, as occurs in the case of *ostensive definition* when one points to something to name it (a process that Augustine alludes to), surely does not solve the issue: pain, memories, purpose, expectations, etc. are examples of words that refer to everyday human experiences that it is quite impossible for one to show (and point to) others. Yet, “[we learn] the concept 'pain' in learning language.” [PI §384] We learn to speak of such concepts like we do for others by *examples* and by *practice* [PI §208]. As Wittgenstein points out, *inner processes* (pain, memories, purpose, expectations, etc.) “[stand] in need of outward criteria” [PI §580] to *become concepts and words* (such criteria are, of course, not necessary *for the inner processes to exist*):

[An] ostensive definition explains the use — the meaning — of a word if the role the word is supposed to play in the language is already clear. [...] We may say: it only makes sense for someone to ask what something is called if he already knows how to make use of the name. [PI §30–31]

Or, in other words:

First, the act of naming presupposes a grammar, or technique of employing a word within a language-game; and second, the mere act of looking inwards does not supply this grammar, or specify a technique of employment. [McGinn, 1997, p. 127]

181 There is, of course, a fuzzy continuum from the general to the particular.

The concepts of “language-game” and “grammar” will be properly defined later. For the time being, suffice it to say that the above is the essence of Wittgenstein's “private language argument”, meant to challenge the belief in an ability of introspection that would enable us to identify and label our sensations – an objection reminiscent of Peirce's critique of introspection, as independently noted by Rosa Calcaterra [2005], who likewise points out that Wittgenstein also argues that truth presupposes the existence of a community (an idea also prominently found in Peirce [CP 5.265]): “we can't talk about ‘right’” if “whatever is going to seem right *to me* is right” [PI §256] (emphasis mine).

Ultimately, Wittgenstein rejects the notion that the meaning of an individual word “is the object for which the word stands” [PI §1], or otherwise that *meaning is correspondence with reality*. Rather, he asserts that the meaning of a word is best thought of as its *use*:

For a *large* class of cases of the employment of the word 'meaning' — though not for *all* — this word can be explained in this way: the meaning of a word is its use in the language. [PI §43]

We find here another element of apparent continuity between the *Tractatus* and the *Investigations*:

If a sign is useless, it is meaningless. That is the point of Occam's maxim. (If everything behaves as if a sign had meaning, then it does have meaning.) [TLP 3.328]

In spite of that similarity, the notion of use in the philosophy of the early Wittgenstein is distinct from that of the later Wittgenstein. For the former, the use of a symbol is a “mode of signification” construed as “syntactic employment” [Johnston, 2007, p. 370]; for the latter, the use of a symbol is the role it plays within a concrete, lived practice of language. Both notions do have in common that meaning can only be understood as part of a context, or as Wittgenstein puts it in the *Investigations*: “Every sign *by itself* seems dead.” [PI §432]

More specifically, the role of symbols in language corresponds to *rules for use* (what Wittgenstein called *grammar*<sup>182</sup>). Accordingly, there is reason to believe that the account given by the later Wittgenstein of meaning could be considered “pragmatic” in a Peircean sense<sup>183</sup>. Peter Skagestad points out a similar isomorphism between the two authors, comparing Wittgenstein's understanding of “I'm expecting him” (when one expects that someone will come) as meaning not a “state of mind” but rather a practical reference to the *possibility* or *potentiality* “I'd be surprised if he didn't come” [PI §577], with Peirce's “view [that] thoughts are inherently dispositional, a view echoing that of Scotus and in turn later echoed by Karl Popper” [Skagestad, 1998].

In light of the above definition of symbols as rules for use, “[the] question 'What is a word really?' is analogous to 'What is a piece in chess?'" [PI §108(d)] This new analogy, which also occurs in the work of de Saussure, among others [Gustafsson, 2014, p. 1], had in fact already come to Wittgenstein's mind as early as 1929:

I can't say: that is a pawn and such and such rules hold for this piece. No, it is the rules which *define* this piece: a pawn *is* the sum of rules for its moves (a square is a piece too), just as in the case of language the rules define the logic of a word. [1929/1975, pp. 327–328 as cited in *Ibid.*, pp. 2–3]

182 While “grammar is *non-autonomous* for the early Wittgenstein [meaning that] to allow for representation, a sign system must be such as to match the logical forms of reality” the later Wittgenstein comes to view it as *autonomous* [Johnston, 2007, p. 389]. Crucially, Wittgenstein remarks that *propositions can have modes of use despite not having any truth conditions*. Wittgenstein has been characterized as an “anti-realist” [See Johnston, 2007], but the philosophical debate as to whether or not he was a “realist” in that specific sense is open [See: <http://www.iep.utm.edu/wittgens/#H7>]. It ought to be noted that the conceptual opposition between “realism” and “anti-realism” is distinct from that between “realism” and “nominalism” previously discussed in 2.1 (c) – Semiotic & pragmatism as interdependent frames of reference, and furthermore that Wittgenstein only *tangentially referred once* to the concept of “realism” in the *Investigations* [PI §402], distinguishing it from “idealism” and “solipsism” but criticizing all three views.

183 In one instance, Wittgenstein explicitly uses “practical consequences” as criterion for meaning: “Why can't my right hand give my left hand money? – My right hand can put it into my left hand. My right hand can write a deed of gift and my left hand a receipt. – But the further *practical consequences* would not be those of a gift.” [PI §268] (emphasis mine)

We find here a theory which appears analogous to the semiotic of Peirce<sup>184</sup> – complementing the analogy drawn between Wittgenstein's notion of meaning as use and Peirce's pragmatism.

An interpretation of Wittgenstein's chess analogy that is consistent with the Peircean theory of general semiotic is that a word (as a *symbol* rather than its instantiation as a *sinsign* or *token*) stands not for a discrete material object, but rather for a rule within a system of such. Indeed, a chess piece need not have any material existence; one may play chess on a computer, or even (and most importantly) through nothing but the pieces' coordinates on an *imaginary* board: “You could imagine [...] having learnt the rules of the game without having ever been shewn [*sic*] an actual piece.” [PI §31 in *Ibid.*, p. 1] A chess game is essentially the representation the participants make of how the rules make up a system.

The rules *qua* representations are nevertheless insufficient to account for the *activity* and *technique* of chess – in the same sense as they cannot account for the technique that “[understanding] a language means to have mastered” [PI §199]. The interpretation of rules (that is, their representation) may not be the only way in which a person can apprehend them:

[There] is an inclination to say: every action according to a rule is an interpretation. But one should speak of interpretation only when one expression of a rule is substituted for another. [PI §201]

A person may alternatively *act either in accord or in conflict with the rule* as it was originally expressed, respectively “what we call 'following the rule' and 'going against it’” [*Ibid.*]. This consists in a *practice*, in the same sense as speaking a language is a practice [PI §197–202].

184 As Joseph M. Ransdell described Peirce's views on the matter: “It is implicit in regarding semiosis as the production of the interpretant by the sign itself that signs are not regarded as being *governed* by rules in the sense of 'falling under' them. The idea is rather that *the disposition or power of the sign to generate an interpretant is the rule*, which thus does not stand over and above the sign, as it were, but is rather an immanent principle therein.” [1992, p. 3] (emphasis mine)

The grammar of the concept of a rule links it, therefore, with a characteristic pattern of activity, or *form of life*, and not with anything that occurs ‘in the minds’ of those who use the rule. [...] Without this surrounding the concept of obeying a rule or going against it is completely empty; for it is nothing that occurs at the time, or ‘in the mind’, that gives the concept of a rule its content, but the existence of a *particular form of practice*. [McGinn, 1997, p. 105] (emphasis mine)

Such a form of practice is called by Wittgenstein a *language-game*. Perhaps the most widely used and debated notion introduced in the *Investigations*, a language-game may be summarized as a vaguely circumscribed assortment of interwoven uses of language-signs, related by *family resemblance*. (“The word ‘language-game’ is used here to emphasize the fact that the *speaking* of language is part of an activity, or of a form of life.” [PI §23]) In like manner, the meaning of a word is construed as a *family of customs* (or one of *habits*, to use terminology borrowed from Peirce [CP 8.317]) delineating the interrelated established uses for it. These customs can be described as the *public rules of grammar*.

Innumerable language-games exist<sup>185</sup>, and “new language-games [constantly] come into existence, [while] others become obsolete and get forgotten” [*Ibid.*]. Wittgenstein also conceives of “the whole, consisting of language and the activities into which it is woven, [as] a ‘language-game’.” [PI 7] However, in spite of his extensive recourse to the concept, Wittgenstein suggested that no definition of a game will ever be wholly satisfactory:

For how is the concept of a game bounded? What still counts as a game, and what no longer does? Can you say where the boundaries are? No. You can *draw* some, for there aren’t any drawn yet. [...]

185 Wittgenstein provides the following list of examples: “Giving orders, and acting on them — Describing an object by its appearance, or by its measurements — Constructing an object from a description (a drawing) — Reporting an event — Speculating about the event — Forming and testing a hypothesis — Presenting the results of an experiment in tables and diagrams — Making up a story; and reading one — Acting in a play — Singing rounds — Guessing riddles — Cracking a joke; telling one — Solving a problem in applied arithmetic — Translating from one language into another — Requesting, thanking, cursing, greeting, praying.” [PI §23]

To repeat, we can draw a boundary – for a special purpose. Does it take this to make the concept usable? Not at all! [PI §68–69]

For the special purposes of this research endeavour, I suggest a simple working definition of *games* as “purposed activity which takes place through signs”<sup>186</sup>. From that perspective, it may not be a stretch to assert that human cognition itself – restrictively defined as an intentional act of interpretation – is a *thought-game*, where the signs through which the game takes place stand for the thoughts themselves.

Such activity is *generally* that of definite rules, but as Wittgenstein emphasizes, seemingly in contradiction with the views of Peirce, it is misleading to think “that if anyone utters a sentence and *means* or *understands* it, he is thereby operating a calculus according to definite rules” [PI §81] Indeed, “is there not also the case where we play, and make up the rules as we go along?” [PI §83] That is to say, “it is only in normal cases that the use of a word is clearly laid out in advance for us” [PI §142]<sup>187</sup>. However, Wittgenstein acknowledges that regularity is essential to language:

[...] if rule became exception, and exception rule; or if both became phenomena of roughly equal frequency — our normal language-games would thereby lose their point. [*Ibid.*]<sup>188</sup>

## **8.8 – Appendix 8: Erving Goffman: *The Presentation of Self in Everyday Life***

The theoretical models of Peirce and Wittgenstein outlined previously are of a philosophical

186 Even this has serious limitations. What about the case of someone speaking out loud while dreaming, or that of a person with Tourette syndrome? Ordinary usage would be to say that such instances are not purposed activity, yet are we to say that they are not language? According to my working definition, they are nonetheless excluded from the category of games, and thus that of language-games.

187 In this regard, on the other hand, Wittgenstein appears to be in accordance with Peirce's view that the meaning of a symbol “inevitably grows, incorporates new elements and throws off old ones” [PPM 25] – introduced in 2.1 (c) – Semiotic & pragmatism as interdependent frames of reference.

188 Likewise, it is necessary that the use of language-signs *mostly* abide by definite rules for the logical and statistical analysis thereof to yield analytically interesting results.

nature, hence they provide somewhat little of use for the more concrete sociological aspects of the analysis I seek to accomplish. To mitigate this shortcoming, I have included a third thinker to complete the theoretical groundwork of my inquiry.

In *The Presentation of Self in Everyday Life* [1959], sociologist Erving Goffman analyzes everyday social interactions as *theatrical performances*. This analogy appears relevant to the study of IRC conversations, inasmuch as they involve people deliberately playing a singular part in structured and controlled social settings that delineate the habitual kinds of expression intended to occur within them (furthermore, fittingly, each utterance is preceded by the name or pseudonym of the speaker, much like a play in the written form).

Goffman's account of social interaction as a performance draws on a number of concepts to elucidate features of phenomena ranging from the nature of identity to that of society. Among the most critical, we find the notion of *social establishment*, which Goffman [1959] defines as:

[...] any place surrounded by fixed barriers to perception in which a particular kind of activity regularly takes place. [...] Within the walls of a social establishment we find a team of performers who co-operate to present to an audience a given definition of the situation. [p. 238]

So defined, social establishments comprise the facts of existence which Goffman suggests to study “from the point of view of impression management” [*Ibid.*], a perspective serving as analytical framework for the analogy of social contact as a theatrical performance. This *dramaturgical approach* is put forward by Goffman as complementary to four alternative, and more common, ways of viewing social establishments (all of which are relevant to my study of IRC) [*Ibid.*, p. 240]:

1. *technical*, consisting in the study of a social establishment as an “intentionally organized system of activity for the achievement of predefined objectives”;

2. *political*: considering the "actions which each participant (or class of participants) can demand of other participants";
3. *structural*: describing the "horizontal and vertical status divisions and [...] kinds of social relations which relate these several groupings to one another";
4. *cultural*: focusing on the "moral values which influence activity in the establishment".

By contrast, the dramaturgical approach will seek to highlight how, in a given social establishment, the performance of some participants serves to convey a particular impression to observers – “a performance presents an idealized view of the situation” [*Ibid.*, p. 35] by way of the previously encountered concept of *definition of the situation*:

The impression and understanding fostered by the performance will tend to saturate the region and time span, so that any individual located in this space-time manifold [can] observe the performance and be guided by the definition of the situation which the performance fosters. [*Ibid.*, p. 106]

In the above quote, speaking of a “space-time manifold” can be misleading: a “region”, if it is defined by the “barriers to perception” which delimit it “to some degree” [*Ibid.*], need not be an *actually tangible place*, and can instead be *potentially tangible*<sup>189</sup>. For that reason, a social establishment may be any social encounter that is set apart from the rest of the world by relatively stable boundaries, and is structured to occur over time with some consistency.

It follows that we may consider an IRC channel, like other online communities, to be a social establishment dedicated to a particular type of collective conversation<sup>190</sup>: it implies both a *region* (fixed barriers to perception) and a *performance* (organized social activity serving to uphold a

189 The lifelong focus of Goffman's research was on face-to-face encounters, but it seems to me that the theoretical models he developed can be applied to all varieties of social interactions.

190 Social establishments resulting from *potentially tangible* arrangements do not exist solely in cyberspace: for example, a collective exchange of traditional paper letters does not quite occur in any *actual* location and time, but nonetheless has a *virtual* structure which mirrors that of an online discussion board.

specific definition of the situation at hand). If we regard the platform as a social establishment, the collective conversations which play out through it can be described and understood by contrast with other social establishments, in accordance with the general features of such facts of existence.

Thus, although the interpersonal interactions defining such an online place are for all intents and purposes only *potentially tangible* (as we cannot, for instance, *actually* hug or kiss someone we meet on IRC whilst this interaction takes place), they nonetheless effectively correspond to permanent restrictions on awareness of what takes places “within”; just as the location of a traditional physical establishment entails that a particular path must be taken to enter, the “place” occupied by an online community in cyberspace coincides with a particular procedure that must be followed to join. Similarly, in both cases, an invitation may or may not be required – on IRC, although most large channels are open for anyone to join, it is possible to create a channel which can be accessed only by a list of invited users.

However, while an IRC channel as just described is like any other social establishment in its most general aspects, in other aspects its features deviate from the norm. This is the case, for instance, with regard to the three functional roles identified by Goffman as crucial to a performance: “those who perform; those performed to; and outsiders who neither perform in the show nor observe it.” [*Ibid.*, p. 144] While the distinction between “outsiders” and “insiders” is rather unambiguous with respect to the community defined by an IRC channel, that between “performers” and their “audience” is fuzzy: there are usually participants who stand out by virtue of their relative liveliness (many are found among administrators and moderators, but not all) and, contrariwise, others who are noticeable owing to their relative silence (in the jargon of those realms, they are known as *idlers* or *lurkers*) – yet

a whole continuum exists between the two poles, and many participants do not stay in a fixed position on this continuum. Goffman acknowledges that “there are [...] times when both functions are performed almost simultaneously by the same others” [*Ibid.*, p. 206], but nowhere does he seem to account for the flexibility of participation found in authentic collective conversations.

Goffman's account of social establishments is only one of two indispensable aspects of his “dramaturgical” conceptual framework as I will hereafter put it into use. The remaining aspect, perhaps of greater importance with regard to my ends, is the sociologist's account of the *self*. The process by which one's individual character develops, described by Goffman as a “machinery of self-production” [*Ibid.*, p. 252], is intrinsically social (thus “the individual is not the natural unit for our consideration but rather the team and its members” [*Ibid.*, p. 149]):

A correctly staged and performed scene leads the audience to impute a self to a performed character, but this imputation—this self—is a product of a scene that comes off, and is not a cause of it. [*Ibid.*, pp. 252—253]

In other words, the self results from conventionalized social activity (which is reminiscent of Peirce's account of the self as a symbol [CP 6.270]: a *virtual entity generated outside of oneself*)<sup>191</sup>, and “the crucial concern, is whether it will be credited or discredited.” [Goffman, 1959, p. 253]

In the same line of thought, Goffman portrays individuals “as *pieces of expressive equipment*, [...] more effective in some ways than non-human parts of the setting” [*Ibid.*, p. 220] (emphasis mine). The notion of expressive equipment, interchangeable with that of *sign-equipment*, describes the symbols used by a performer as part of a scene. If the performer *qua* individual is itself a piece of expressive equipment, this serves to confirm that Goffman's micro-sociological analysis is akin to

191 A notion introduced in 1. 2 – General Object of Inquiry and Aim of this Research.

Peirce's approach with regard to *the essentially symbolic nature of the self*. For Goffman, this symbolic self is an idealized character ascribed to a performer as a corollary of impressions conveyed (more or less correctly) by the performer, through the use of the limited expressive equipment at his disposal, to preserve the definition of the situation. In identity matters, "the idea to which [the sign] gives rise" [CP 1.339] is the *interpretant* [*Ibid.*] as outcome of the social interaction.

Goffman calls *front* "the expressive equipment of a standard kind intentionally or unwittingly employed by the individual during his performance" [1959, p. 22], usually already present before a particular performance begins [*Ibid.*, p. 27] (a same front will generally be used for many performances). It may be divided into the *setting*, "involving furniture, decor, physical layout, and other background items" [*Ibid.*], and the *personal front*, which comprises for instance "insignia of office or rank; clothing; sex, age, and racial characteristics; size and looks; posture; speech patterns; facial expressions; bodily gestures; and the like" [*Ibid.*, p. 24], and the latter may itself be divided into *appearance*, which "[tells] us of the performer's social statuses [and] temporary ritual state" [*Ibid.*] (e.g. work, formal social contact, leisure...), and *manner*, which "[warns] us of the interaction role the performer [expects] to play" [*Ibid.*] (e.g. dominating, peaceful, deferential...)<sup>192</sup>.

We find that a performer, making use of expressive equipment, will typically engage in two correlative practices to maintain the "expressive coherence" [*Ibid.*, p. 63] of a situation. In another book, Goffman writes: "Given [one's] attributes and the conventionalized nature of the encounter, [one] will find a small choice of lines will be open [...] and a small choice of faces will be waiting

192 As Goffman appears to assert that individuals *per se* are part of the *setting*, we may wonder if this should be taken to mean that the part they play is analogous to that of furniture: elements of the background, while it is rather their *personal front* which supplies the foreground of their performance.

[...]" [1967, p. 7] The concept of *face* refers to what a performer has or maintains when the impressions he conveys are internally consistent and in agreement with the messages conveyed by other expressive equipment (human or not) in the scene [See *Ibid.*, p. 6]; hence, the two interrelated means of preserving expressive coherence can equally be construed as means of *maintaining face*.

Firstly, a performer will "conceal or underplay those activities, facts, and motives which are incompatible with an idealized version of himself and his products" [Goffman, 1959, p. 48], and inversely will use sign-equipment to express a "favorable social style" [*Ibid.*, p. 36]. "For if the individual's activity is to become significant to others, he must mobilize his activity so that it will express during the interaction what he wishes to convey." [*Ibid.*, p. 30] Presumably, typical online communities empower performers to accentuate the idealized nature of their performance by largely filtering out involuntary communication; vice versa, it is a well-known fact that their very nature facilitates pretense.

Secondly, "[...] the performer [will segregate] his audiences so that the individuals who witness him in one of his roles will not be the individuals who witness him in another of his roles." [*Ibid.*, p. 137] Roles appear when people interiorize, and subsequently put into practice, the rules of conduct that they hold to correspond to the idea they have of themselves as part of each performance in which they take part (to reformulate, we might say that a role is the result of *conventionalized self-representation*). In ordinary social establishments, performers will adhere to relatively definite interaction roles in order to uphold the definition of the situation. This applies equally to physical and online communities: in both cases, an individual will customarily act in accordance to different roles when taking part in different performances, and different performances will correspond to different

audiences. However, off-line and online alike, we will often find that distinct performances overlap to varying extents. Similarly, different roles involve rules of conduct that share commonalities.

Finally, the two aforementioned practices, being rules of conduct in and of themselves, frequently overlap. This is perhaps most obvious when a performer seeks to maintain a particular idealized definition of himself by refusing to disclose information pertaining to a performance other than the one immediately at hand (for example, many users of some IRC channels will be uncomfortable disclosing aspects of their off-line life).

At first glance, Goffman's account of social interaction may appear dated for a study of cyberspace, as it was formulated long before its advent; the examples that the sociologist had in mind as he developed his theory were well-circumscribed material places. At second glance, it occurs to me that this may make the theory *more* valuable for my purposes rather than *less*: contrasting the reality of digital realms with a conceptual framework developed in complete ignorance of those realms should highlight how they compare to traditional settings that served, and presumably still serve, similar purposes. Although the same assessment applies to the theories of Peirce and Wittgenstein as well, their abstract and general nature entails that the two philosophers spoke comparatively little of matters that immediately depended upon the thinkers' social and historical context.

## **8. 9 – Appendix 9: Additional Methodological Concerns**

### **8. 9 (a) – Content Analysis & Computational Linguistics**

The statistical methods put into practice for the purposes of this thesis may be construed as being conceptually situated at the intersection of two well-established methodological frameworks for the analysis of text data: *content analysis* (primarily *quantitative content analysis*, although other forms are also relevant to a lesser extent) and *computational linguistics*, both of which ostensibly apply to the present study in some respects.

In the most general sense, content analysis is first and foremost a scientific method for the study of communication through the examination of its products. In a more specific sense, however, various operational definitions of content analysis compete in the literature, which primarily depend on the associated epistemological frameworks (and corresponding analytical purposes). Nonetheless, such definitions tend to include the notion of a methodical and replicable analysis of the informational content generated in the course of communication, from which logically valid inferences can be made about underlying abstract phenomena [Krippendorff, 2004; Riffe et al., 2005 (2008)]: that is to say, content analysis fundamentally attempts to draw a causal link between a theoretical phenomenon that cannot itself be observed (for example, semiosis) and its material outcome in the form of a text. Indeed, it is necessarily only through analysis of the latter that it is possible to comprehend the natural processes by which it came into existence.

In addition, content analysis is usually thought to involve a systematic protocol for data processing – including an explicit set of rules, such as an algorithm used to compute important statistical measures – to ensure that results can be replicated and provide useful insights for further

inquiries. Indeed, the potential of content analysis to involve a standardized set of rules for the systematic description and reliable interpretation of meaningful data resulting from communication processes is, above all, what makes it uniquely valuable for scientific research.

Content analysis is extremely versatile and the various forms it can take can therefore be useful in all fields of social sciences, whether to understand the process of communication itself or to make sense of communication within the actual contexts in which it occurs (for instance, the rise of particular social phenomena). In all cases and regardless of the precise form it takes, however, through the analysis of observable signification-bearing content, this type of method may provide empirical evidence to support, or conversely to challenge, theory-derived hypotheses about communication and associated human behaviours.

For its part, the similar framework of computational linguistics is generally defined as the use of machines to analyze natural language, typically by having recourse to sophisticated computer programs (algorithms) that aim to uncover hidden patterns in language. In other words, according to the *Association for Computational Linguistics*:

“Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical").<sup>193</sup>

Computational linguistics is in fact an interdisciplinary field, which draws on theories from linguistics for the development of statistical procedures and formal, computational models to

193 “What is Computational Linguistics?” section on the page “What is the ACL and what is Computational Linguistics?” of the official website of the *Association for Computational Linguistics*. – <https://www.aclweb.org/portal/what-is-cl>

understand the logical structure and generation of language [Mitkov, 2003 (2005)]. The practical applications for such procedures and models are often described within the context of artificial intelligence, given that one of the field's primary areas of focus is the computer modelling of actual instances of human language use [Uszkoreit, 1996 (2000)]. Notwithstanding, the present thesis incorporates aspects of computational linguistics through the use of computer-based statistical analyses to highlight linguistic patterns within the communication processes that take place on IRC. Moreover, its use of diagrammatic representations is specially hypothesized to have the potential to reveal complex structural features of human sign-generation within on-line communities.

Whether they are put to use according to the methodological framework of content analysis or according to that of computational linguistics, sophisticated computational tools can facilitate a comprehensive examination of the patterns of language use that exist across the constantly expanding collections of data available in cyberspace. In particular, visualization tools like KH Coder – the software used in this work<sup>194</sup> – may help to get an intuitive sense of the overreaching signification of vast amounts of textual content through the computational production of diagrams.

That being said, along with computer-enabled statistical analysis, human interpretation of its results – with respect both to mathematical features and to verbal content – remains crucial to reach an adequate understanding of the meaning of data in relation to its context.

“It follows that content analysts must look outside the physicality of texts – for example, to how people other than the analysts use these texts, [or] the conceptions and actions the texts encourage. This requirement is a key to understanding the limitations inherent in computer text analysis. [...] Without human intelligence and the human ability to read and draw inferences from texts, computer text analysis [...] [operates] in the contexts of [its] users' worlds without understanding those contexts.”

194 As detailed in the section 3. 2 – Tools Used.

---

[Krippendorff, 2004, p. 23]

An important implication of the above is that, in the process of analyzing textual content, a researcher cannot but draw conclusions within a predefined context. Furthermore, how the same context is construed can differ greatly from one scientific discipline to another, and no single interpretation is the only valid one. The same data-set can be adequately interpreted from many angles, depending on the analyst's purposes. Thus, it is important to explicitly define one's theoretical assumptions, as well as one's research question, which will both guide one's interpretation of all results obtained: any meaningful conclusion pertaining to text data is bound by the analyst's purpose and interpretative framework.

Ultimately, this research endeavour aims to overcome the practical distinction between qualitative and quantitative approaches to the analysis of text data – notably by integrating within the same section, for each data-set, an enumeration of prominent statistical properties along with an effort to ascribe a plausible meaning to these mathematical features (as well as plausible ways in which these features might be, in turn, conceptually linked with each other). Given how the production of diagrams summarizing text data involves, in a sense, automatically *reading* this data, and that interpreting such diagrams also calls for me to *read* them, the complex nature of the methods used may be better understood in light of this remark: “Reading is fundamentally a qualitative process, even when it results in numerical accounts.” [*Ibid.*, p. 20]

#### 8. 8 (b) – Methodological pansemioticism:

I hold the aforementioned conception of cognition/communication to be conducive to the

adoption of what might be termed *methodological pansemioticism*<sup>195</sup> for the study of cyberspace. First and foremost, this entails a heuristic emphasis on the omnipresence of signs in the course of computer use: every computer employed to access the Internet becomes a purposed instantiation of the signifying potential of the *virtually infinite* network of signs which makes up cyberspace.

Thus, digital media are viewed primarily as means to extend the capabilities of the human mind, in the same sense as other technologies of information and communication increase the range of possible signifying processes. With regard to cyberspace, this is to be understood as the inevitable incorporation of *computer algorithms* into human cognition. It seems very likely that Peirce foresaw the prospect of further augmentation of human cognition through technology, considering how he pointed out that “[one’s] faculty of discussion is [...] localized in [one’s] inkstand” [CP 7.366] in precisely the same sense as it is “localized” in a part of one’s brain, and that he even called attention to the fact that “the mind working with a pencil and plenty of paper” [W 6:70-71] has less limitations than the “unaided mind” [*Ibid.*]... what is more, in an article about *Logical Machines*!

Peirce probably did not foresee, as did Vannevar Bush, that there would one day be an “enlarged intimate supplement to [one’s] memory” [1945] that would enable the *near-instantaneous consultation of the “common record”* [*Ibid.*] *of a sizable portion of the human kind*. However, we may think that to describe the semiotic processes made possible by the advent of the Internet, Peirce might have said that *our faculty of thinking is now also in our computers, and in cyberspace as well*.

#### 8. 9 (c) – Similarity or dissimilarity of channels studied

<sup>195</sup> In its most general sense, *pansemioticism* is the view that everything is (actually or potentially) a sign.

As we will remember from [3. 4 – Sample Selection](#), a major limitation of case studies is the tendency of researchers to consider only a few selected cases that are likely to confirm their theoretical leanings. To what extent are the channels selected similar or dissimilar? What are the common threads – to what users purport to do in them – that would explain the law-like regularities (and apparent teleological mechanism) observed here?

It might be argued that most of the channels studied are “serious” in nature, having a clearly defined purpose related to a “practical” goal. However, the #ubuntu-offtopic channel provides an example of a different nature, being dedicated to casual chit-chat between like-minded people (users of the Ubuntu operating system). On the other hand, all channels studied except one (#politics) are official channels of non-profit initiatives, while one large group and one sub-group of channels that were part of my analysis had to do with non-profit, Internet-based collaborative projects (Ubuntu and Wikimedia). Channels associated to non-profit initiatives, in general, have the advantage of being very open to the general public, and they often make their logs publicly available, thus precluding most ethical issues, and many practical issues, that the researcher may face in the study of Internet-mediated communication [See Uthus & Aha, 2013]. In the latter regard, I can only lament the limitations of my own resources.

Another major limitation comes from the fact that the data analyzed were overwhelmingly collected from English-language channels, and the results of the analysis were entirely in English.

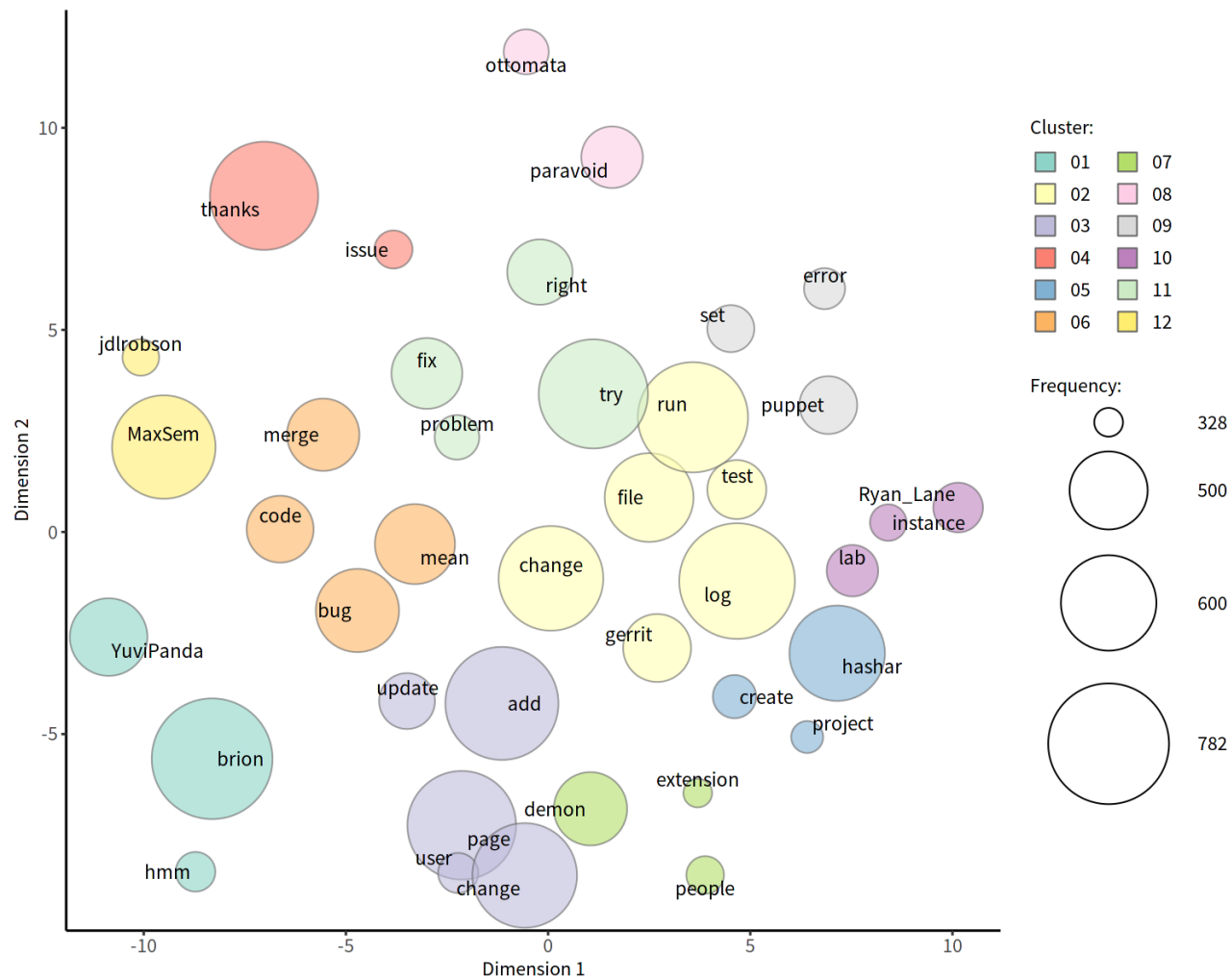
On the other hand, one of the most analytically powerful characteristics of the data-sets selected is that they are international in nature. This largely avoids the bias of ethnocentrism that could otherwise apply to the study of the uses of IRC [See Pastinelli, 1999, 2002], and also allows for

---

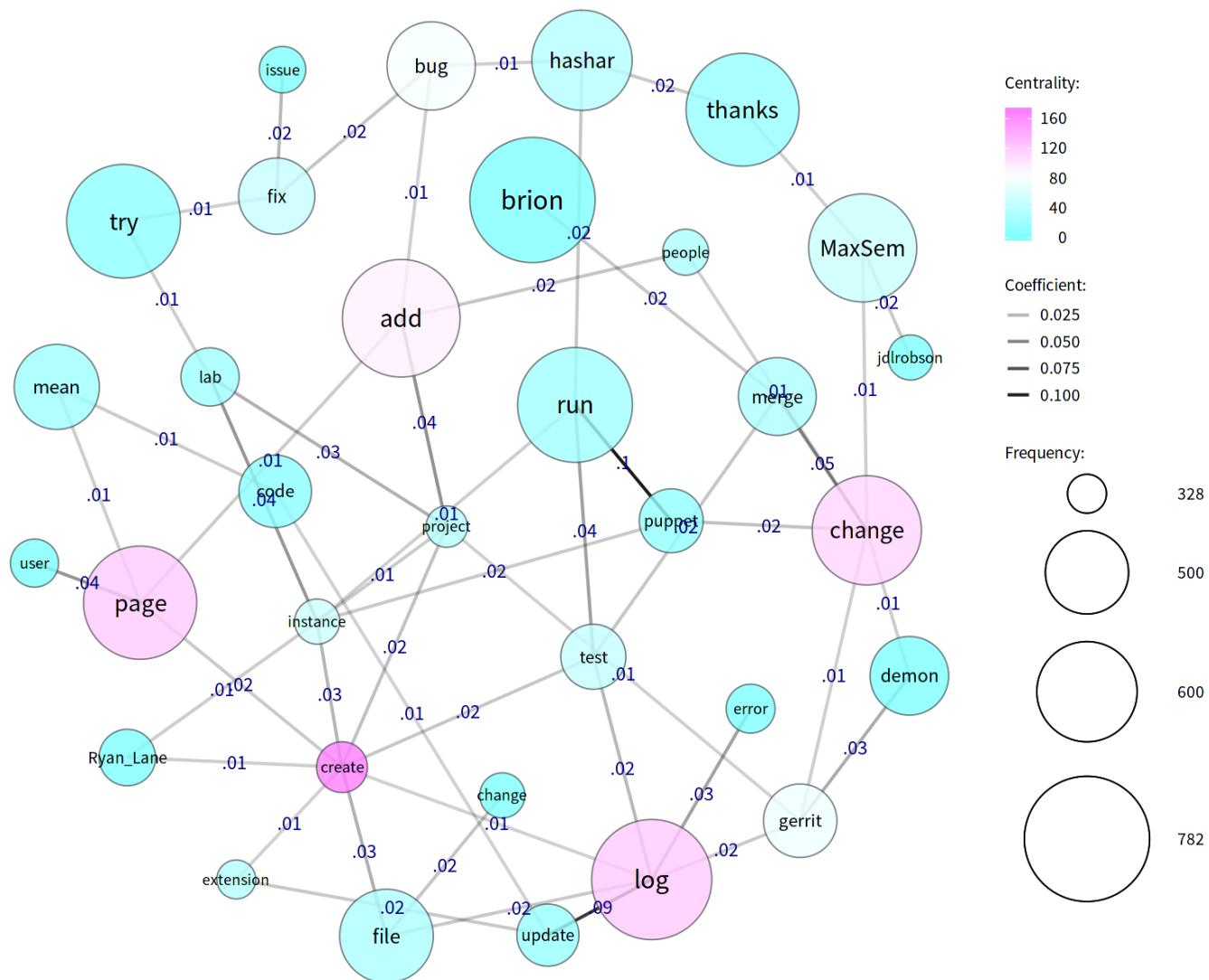
a more accurate representation of the specificity of online communities that may appear on IRC: the heterogeneity of geographic locations of participants within a same virtual “space” is one of the most prominent traits of cyberspace.

## 8. 10 – Appendix 10: Additional Diagrams – Wikimedia-related channels (KH Coder, Part 3)

Figure 7a: “Multi-Dimensional Scaling” diagram representing the #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office and #wikimedia-operations IRC channels in June of 2013 (words = 40, clusters = 12).



**Figure 7b:** “Co-Occurrence Network of Words” diagram representing the #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office and #wikimedia-operations IRC channels in June of 2013 (words = 40; edges = 50).



8. 10 (a) – Interpretation of *Figure 7a* and *Figure 7b*:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“brion”, “log”, “add”, “run”, “try”, “page”, “thanks”, “change”, “MaxSem”, “hashar”, “file”, “bug”, “mean”, “YuviPanda”, “demon”

Among the word-stems above, “try” is notable in that it is the only one that is also part of the analogous list for every single data-set analyzed until now. However, it is here in the 5<sup>th</sup> position, by contrast with the three Ubuntu-related data-sets apart from *#ubuntu-offtopic* (2014-08-08 to 2014-10-13)-, where it is systematically the 2<sup>nd</sup> element in the list – this probably echoes how the pertinent Ubuntu-related channels, as opposed to the Wikimedia-related channels now under consideration, are mainly dedicated to technical assistance and thus prominently involve a process of *trial and error*. On the contrary, something that the same three Ubuntu-related channels have in common with this one is the inclusion of “thanks” among their most frequent signs, which suggests that the primarily practical purpose and more or less technical nature of these channels is reflected in how the participants often express their gratitude towards the contributions of others.

We also find the word-stem “mean” among the most frequent in the majority of other data-sets previously discussed, except ‘all Ubuntu-related channels’ (2004-12-25) and *#kubuntu* (2013-01-01 to 2013-01-31), but its relative location in the list above is most similar to where we find it in the case of the ‘English-language Ubuntu-related channels’ (2014-07-01 to 2014-08-31) data-set. The meaning of this similarity is unclear, unlike that of the presence of “add” and “bug” exclusively in the latter data-set and this one: these word-stems can easily be interpreted in light of their possible uses in computing, such as

how an Ubuntu user might *add* new sources of software or file a *bug* report online, and how a Wikimedia Foundation contributor might *add* a new page to one of the Foundation's websites or fix a *bug* in a software project maintained by the organization. By the same rationale, it is not at all difficult to think of plausible reasons – many examples of which have been given in previous sections – for the word-stem “run” listed above being also one of the most common in all four Ubuntu-related data-sets, and similarly for the word-stem “file” but without the *#ubuntu-offtopic* (2014-08-08 to 2014-10-13) data-set. All things considered, with the only word-stems that appear to be uniquely listed among the most frequent ones in this particular data-set being pseudonyms along with “log”, “page”, and “change”, it is worth highlighting that a basic frequency analysis revealed a surprising level of resemblance between this data-set and Ubuntu-related ones.

Leaving aside other data-sets, a comparative overview of [Figures 8a](#) and [8b](#) shows that *cluster 03* includes, among the 5 word-stems that appear as most central in [Figure 8b](#), 3 which are also among the most frequently occurring: “add”, “page”, and “change” – in addition to the much less frequent word-stems “update” and “user”. Given that all the channels under consideration globally relate to the *Wikimedia Foundation*, which manages several websites including *Wikipedia*, the term “page” here surely refers mainly to web *pages*, several of which were probably modified on a regular basis, as highlighted by the inclusion of word-stem “change” next to “page”. Therefore, cluster 03 in its entirety refer to conversations about *updating*, *changing* or *adding* web *pages*, such as *users'* personal pages, as reflected by the strong word-association “user—page” ( $C = 0,04$ ).

It seems accurate to say that *cluster 02* is the central cluster in [Figure 8a](#), and accordingly, the elements grouped together within it – “log”, “run”, “file”, “change”, “gerrit”, and “test” -- are

specially useful to understand the common thread of signification running through most conversations in Wikimedia-related channels during the period analyzed. The least ambiguous part of the list is ostensibly “gerrit”, which corresponds to the *Gerrit* software<sup>196</sup> – described by Wikipedia as “a free, web-based team code collaboration tool”<sup>197</sup> – which is extensively used for the Wikimedia Foundation’s software development projects<sup>198</sup>.

Otherwise, this cluster includes a second instance of the word-stem “change”, which necessarily corresponds to a rather different pattern of use compared to the other instance in *cluster 03* (described above)<sup>199</sup>, along with the word-stems “log” and “file” – the later being used in the usual way relevant to computing and the former being used either as a verb synonymous with “to record” or to describe the resulting data – and finally the word-stems “run” and “test”, which simply mean that the users of these channels frequently *run* various *tests*, often to make sure the websites maintained by the Foundation function as intended, an interpretation that is highlighted by the strong word-association “run—test” ( $C = 0,04$ ). A conceptual link between *cluster 02* and *cluster 03* is found in the word-association “update—log” ( $C = 0,09$ ), which describes the procedure by which one makes sure that the most recent data available is copied into the corresponding log.

The meaning of the elements in *cluster 02* may be better understood in light of *cluster 09*, which contains “puppet”, “error”, and “set”. Here, in fact, “puppet” refers to the *Puppet* software<sup>200</sup>, which the relevant Wikipedia page defines as “an open-source software configuration management

196 *Gerrit Code Review*’s official website: <https://www.gerritcodereview.com/>.

197 See the Wikipedia page: [https://en.wikipedia.org/wiki/Gerrit\\_\(software\)](https://en.wikipedia.org/wiki/Gerrit_(software)).

198 The *Foundation* runs its own instance of Gerrit at the following URL: <https://gerrit.wikimedia.org>.

199 Like similar “duplicates” encountered in other diagrams, one is perhaps generally used as a noun, and the other as a verb.

200 Official website of *Puppet*: <https://puppet.com/>. It is worth noting that the Wikimedia Foundation maintains its own version of this software, which can be found on GitHub: <https://github.com/wikimedia/puppet>.

tool”<sup>201</sup>. The word-association “run—puppet” ( $C = 0,1$ ), the strongest in its graph, emphasizes the importance of *running* the *Puppet* software to manage the configuration of the Foundation’s web servers. The word-stem “error” included in the same cluster as “puppet” might reflect that the software regularly experiences issues, while the moderately strong word-association “log—error” ( $C = 0,04$ ) plausibly indicates that *error* messages requiring attention are recorded in the project *log* – or, alternatively, that the process of recording a given log encountered an error. The last among the three word-stems in *cluster 09*, “set”, is hardly enlightening in this case, but its proximity to “puppet” might mean that some parameters of the *Puppet* software are what is being *set*, although this association does not seem very systematic given the absence of any corresponding word-association in [Figure 8b](#).

*Cluster 07*, which contains the signs “bug”, “code”, “merge”, and “mean”, seems to mainly feature a more technical description of the process of resolving *bugs* in computer *code*. The word-stem “merge” is a common technical term in programming referring to the operation of integrating changes made to a computer program (that is, to code) into the working version, usually by uploading them to a collaborative “version control” platform such as *GitHub*. This interpretation is supported by the very strong word-association “merge—change” ( $C = 0,05$ ), among others, in [Figure 8b](#). For its part, the word-stem “mean”, which can once more be interpreted in many vastly different ways, might either illustrate that a user *means* to apply some changes to a web page, or refer to the *meaning* of such changes.

*Cluster 04* includes the word-stems “try”, “problem”, “fix”, and “right”. While the word-stem

201 Quote taken from the Wikipedia page: [https://en.wikipedia.org/wiki/Puppet\\_\(software\)](https://en.wikipedia.org/wiki/Puppet_(software)).

“right” might indicate that participants on the channels often find themselves in agreement with each other, the rest of the cluster ostensibly refers to exchanges about *problems* that need to be *fixed*, and the process of *trial-and-error* that is often required to find solutions in such contexts. The word-stem “issue”, in *cluster 11*, is located close to “problem” and likely more or less always used as a synonym.

The remainder of *cluster 11* is only the word-stem “thanks”, located at the periphery in a corner of the diagram, probably often used in a distinctive fashion to express one’s gratitude when issues are solved. In a similar way, *cluster 01* encloses the sign “hmm”, which is an onomatopoeia typically “used to express the action or process of thinking”<sup>202</sup>, and which has a pattern of use rather different from that of ordinary words as shown by its position in the corner of the diagram.

*Cluster 05* includes the three word-stems “hashar”, “create”, and “project”. The first element, “hashar”<sup>203</sup>, is the pseudonym used by a Wikimedia Foundation employee, while the other two elements likely refer to the *creation* of Web *projects* – despite the word-association “create—project” ( $C = 0,02$ ) being relatively weak in the other diagram, equally so as “create—page” ( $C = 0,02$ ). There does exist, nevertheless, a moderately strong association between the word-stems “create—file” ( $C = 0,03$ ), as well as between the word-stems “create—instance” ( $C = 0,03$ ), hinting at the fact that in this data-set, the concept of *creating* is typically used in an impersonal, technical sense to strictly describe operations performed on a computer, rather than anything pertaining to what we might usually call “creativity”.

*Cluster 10*, finally, is composed of the three signs “Ryan\_Lane”, “instance”, and “lab”,

202 Definition provided by the Merriam-Webster online dictionary:

<https://www.merriam-webster.com/dictionary/hmm>.

203 Antoine Musso, a software engineer for the Wikimedia Foundation and developer of the MediaWiki software.

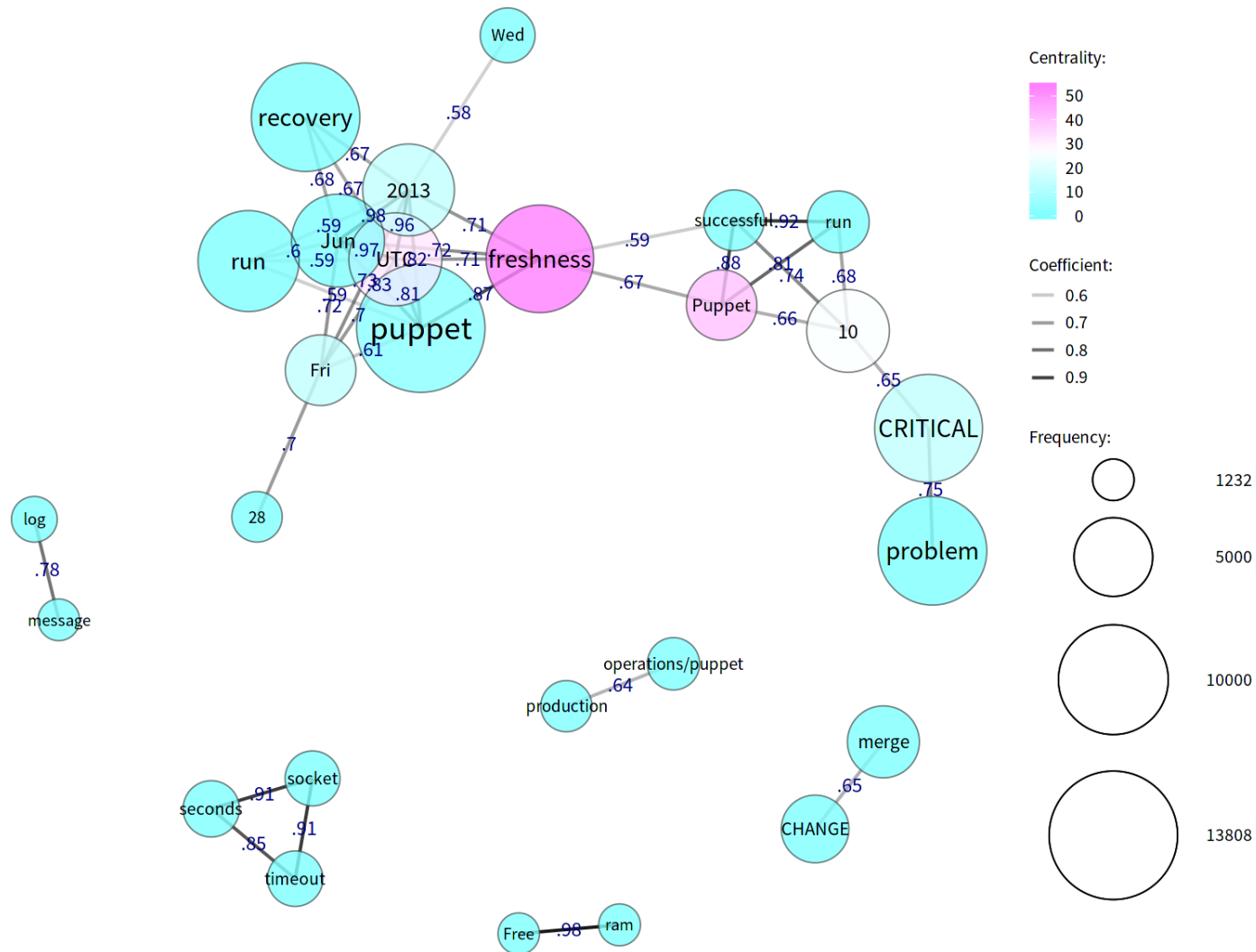
[\[http://www.mediawiki.org/wiki/User:Hashar\]](http://www.mediawiki.org/wiki/User:Hashar)

pertains to a Wikimedia Foundation volunteer, *Ryan Lane*<sup>204</sup>, who happens to have been responsible for the creation of *Wikimedia Labs*, which “provides cloud computing infrastructure”<sup>205</sup> – thus, the strong word-association “lab—instance” (C = 0,04) seemingly corresponds to executing *instances* of processes related to Wikimedia Labs.

204 Ryan Lane’s user page on the MediaWiki website: [https://www.mediawiki.org/wiki/User:Ryan\\_lane](https://www.mediawiki.org/wiki/User:Ryan_lane).

205 See the page: [https://www.mediawiki.org/wiki/Wikimedia\\_Labs](https://www.mediawiki.org/wiki/Wikimedia_Labs).

**Figure 7c:** “Co-Occurrence Network of Words” diagram representing the #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office and #wikimedia-operations IRC channels in June of 2013, exceptionally including all lines by known “bots” (words = 40; edges = 40).



### 8. 9 (b) – Interpretation of Figure 7c:

For the sake of consistency, the 15 most frequently occurring word-stems are listed below (in decreasing order of frequency):

“puppet”, “master”, “problem”, “recovery”, “CRITICAL”, “freshness”, “run”, “UTC”, “output”, “Jun”, “Patchset”, “review”, “merge”, “Fri”, “Puppet”

The above has relatively little in common with the corresponding lists for other data-sets previously analyzed – indeed, an abnormal proportion of the list’s elements are found in no other such list (excluding the other two that also include “bots”<sup>206</sup>): “master”, “recovery”, “CRITICAL”, “freshness”, “UTC”, “output”, “Jun”, “Patchset”, “review”, “merge” “Co-Occurrence Network of Words” , and “Fri”. Few words that are ever used in ordinary English conversations are included, and even those are evidently used in a way quite different from their meaning in everyday conversations.

In the context of IRC, “bots”<sup>207</sup> are relatively simple computer programs, designed to fulfill a particular function for which they are required to superficially mimic IRC users from the perspective of the IRC protocol. In some IRC conversations, unless they are deleted from them, their extensive involvement clutters the corresponding diagrams – especially “Co-Occurrence Networks of Words”, for reasons that will be detailed later – with redundant statements, derived from rigid templates, that are automatically generated as a consequence of predefined triggers (such as the time of the day, or the writing of a specific message in the channel).

As is evidently the case in [Figure 7c](#), the inclusion of “bots” tends to result in exceptionally close-knit and geometrically regular networks of signs, involving word-associations that are extremely

<sup>206</sup> [Figure 8c](#) and [Figure 9c](#).

<sup>207</sup> In Ubuntu-related channels, for example: “One of these bots, ubottu, contains a collection of factoids (short messages), which can be used to answer other people's questions.” [Uthus & Aha, 2013]

strong, reflecting the fact that the “habits” of a simple computer program are *absolutely rigid*. As a matter of fact, the coefficients in this diagram are all far higher (the legend ranges from 0.6 to 0.9, while the highest number included in the corresponding legend of a diagram excluding bots is 0.125, in [Figure 5b](#)) than in any other “Co-Occurrence Network of Words” diagram in this thesis except for [Figure 8c](#) and [Figure 9c](#), which also include “bots”.

Indeed, the fact that the language automatically generated by “bots” features the same combinations of words over and over again is why this type of diagram is much more sensitive to the inclusion of these lines: a “Co-Occurrence Network of Words” diagram includes only the word-stems that are at once among the most frequent and *also* part of the strongest word-associations. When lines by “bots” make up a large proportion of the data-set, the strongest word-associations involve only a small number of interconnected elements – in fact, in [Figure 7c](#) and [Figure 8c](#), such a small number of elements are mutually linked that only 40 edges could be represented in the graphs instead of the usual 50. Any additional edges end up being superimposed over others, thus making parts of the diagram illegible.

It was deemed unnecessary to generate “Multi-Dimensional Scaling” diagrams for data including bots, given how “Co-Occurrence Network of Words” diagrams are much more representative of the distinctive patterns that characterize such texts. In [Figure 7c](#), for example, the oddly shaped largest and topmost cluster, including the most central elements “freshness”, “puppet”, and “UTC”, represents lines such as those in the excerpt below:

```
<icinga-wm>    RECOVERY - Puppet freshness on tola is OK: puppet ran at Sun Jun 9
04:31:24 UTC 2013
```

```
<icinga-wm>    PROBLEM - Parsoid on wtp1022 is CRITICAL: CRITICAL - Socket timeout
```

after 10 seconds

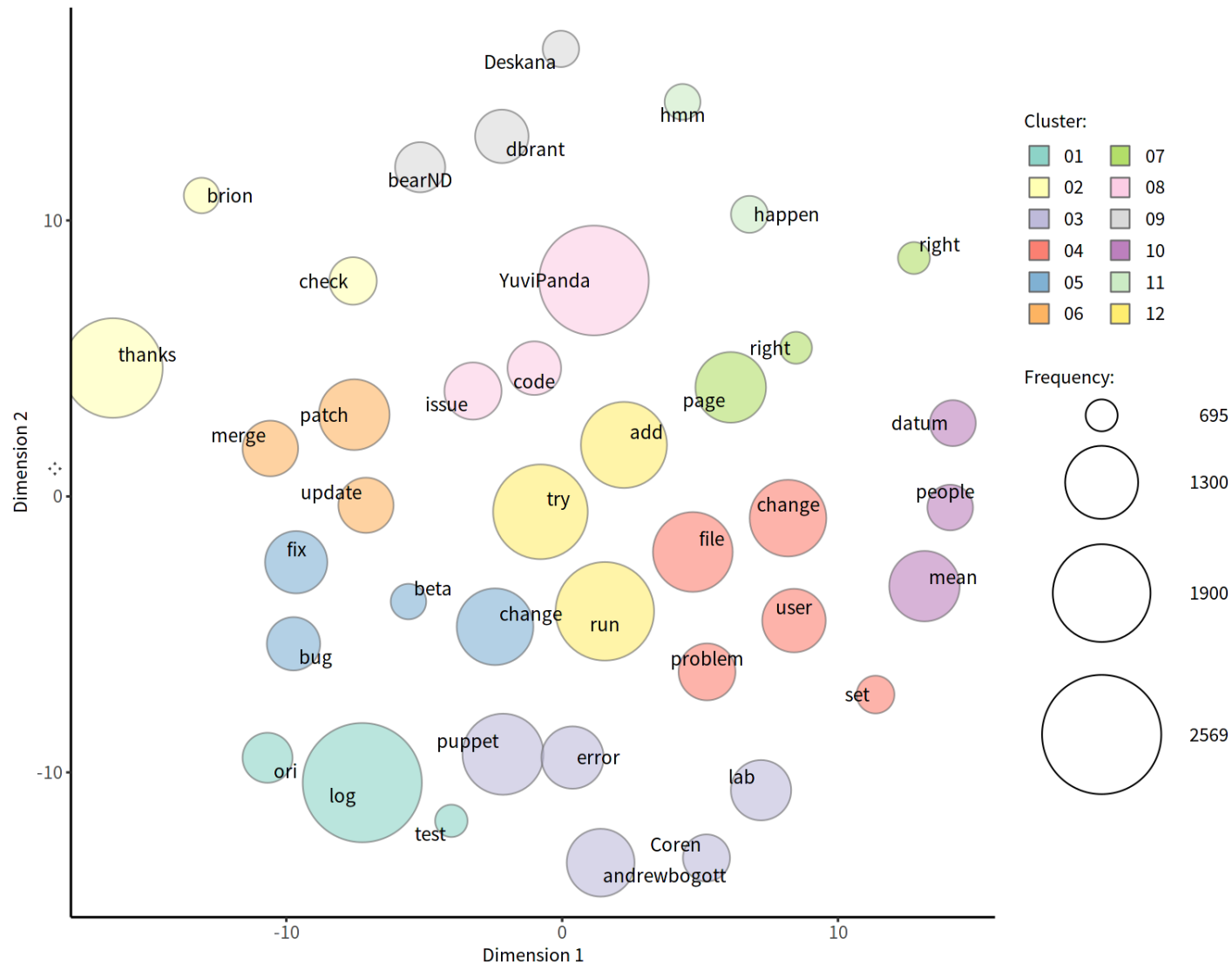
<icinga-wm> PROBLEM - Puppet freshness on tola is CRITICAL: No successful Puppet run in the last 10 hours

<icinga-wm> RECOVERY - Puppet freshness on kuo is OK: puppet ran at Sun Jun 9 04:32:45 UTC 2013

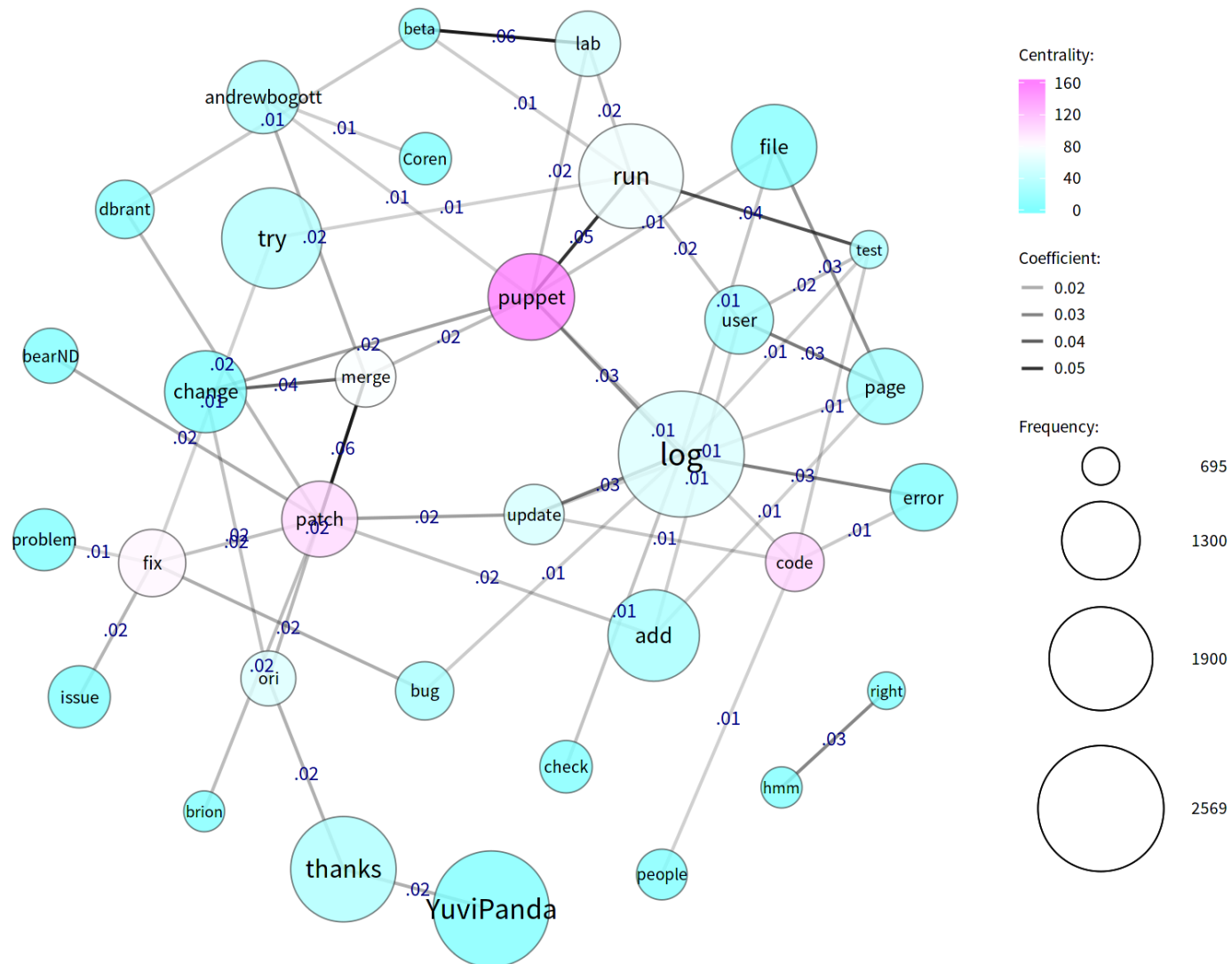
<icinga-wm> PROBLEM - Puppet freshness on kuo is CRITICAL: No successful Puppet run in the last 10 hours

The lines above notably illustrate how the word-associations “Jun—2013” ( $C = 0,98$ ), “UTC—Jun” ( $C = 0,97$ ), and “UTC—2013” ( $C = 0,96$ ) simply correspond to the date, while other word-associations that are included in that excerpt include most of the remaining strongest ones.

**Figure 8a:** “Multi-Dimensional Scaling” diagram representing all publicly logged IRC channels managed by the Wikimedia Foundation except #wikimedia-dev in July and August of 2014 (words = 40, clusters = 12).



**Figure 8b:** “Co-Occurrence Network of Words” diagram representing all publicly logged IRC channels managed by the Wikimedia Foundation except #wikimedia-dev in July and August of 2014 (words = 40; edges = 50).



### 8. 9 (c) – Preliminary interpretation of *Figure 8a* and *Figure 8b*:

Again, it is useful to begin with a list of the 15 most frequently occurring word-stems, which are listed below (in decreasing order of frequency):

“log”, “YuviPanda”, “thanks”, “run”, “try”, “add”, “puppet”, “file”, “change”, “right”, “patch”, “page”, “mean”, “andrewbogott”, “user”

In the list of word-stems above, “try” is included yet again at the 5<sup>th</sup> position. In addition, the word-stem “thanks” is once more part of this list, most likely for the reasons stated in the preceding section. Likewise, we can again note the presence of the word-stems “run”, “add”, “file”, and “mean”. As for the elements of this list previously identified as defining features of the *Wikimedia-related* data-sets, we can still observe the presence of pseudonyms, as well as the word-stems “log”, “change”, and “page”. However, in contrast with the previous data-set, the word-stems “puppet”, “right”, “patch”, and “user” are featured instead of the word-stem “bug”, and comparatively less pseudonyms are included.

With respect to more detailed analyses, the specific word-stems contained in the figures for this section and the way in which they relate to each other are essentially similar to the corresponding figures of the preceding section. However, as the structure of those two sets of diagrams does differ in important respects, this aspect will be the one considered for the remaining examination of this section’s channels.

Looking at the general configuration of Figure 8a, one striking aspect is the absence of a well-defined central cluster. It even seems relatively difficult to distinguish meaningful individual clusters. As a matter of fact, the data illustrated in that figure appears to be fairly uniformly

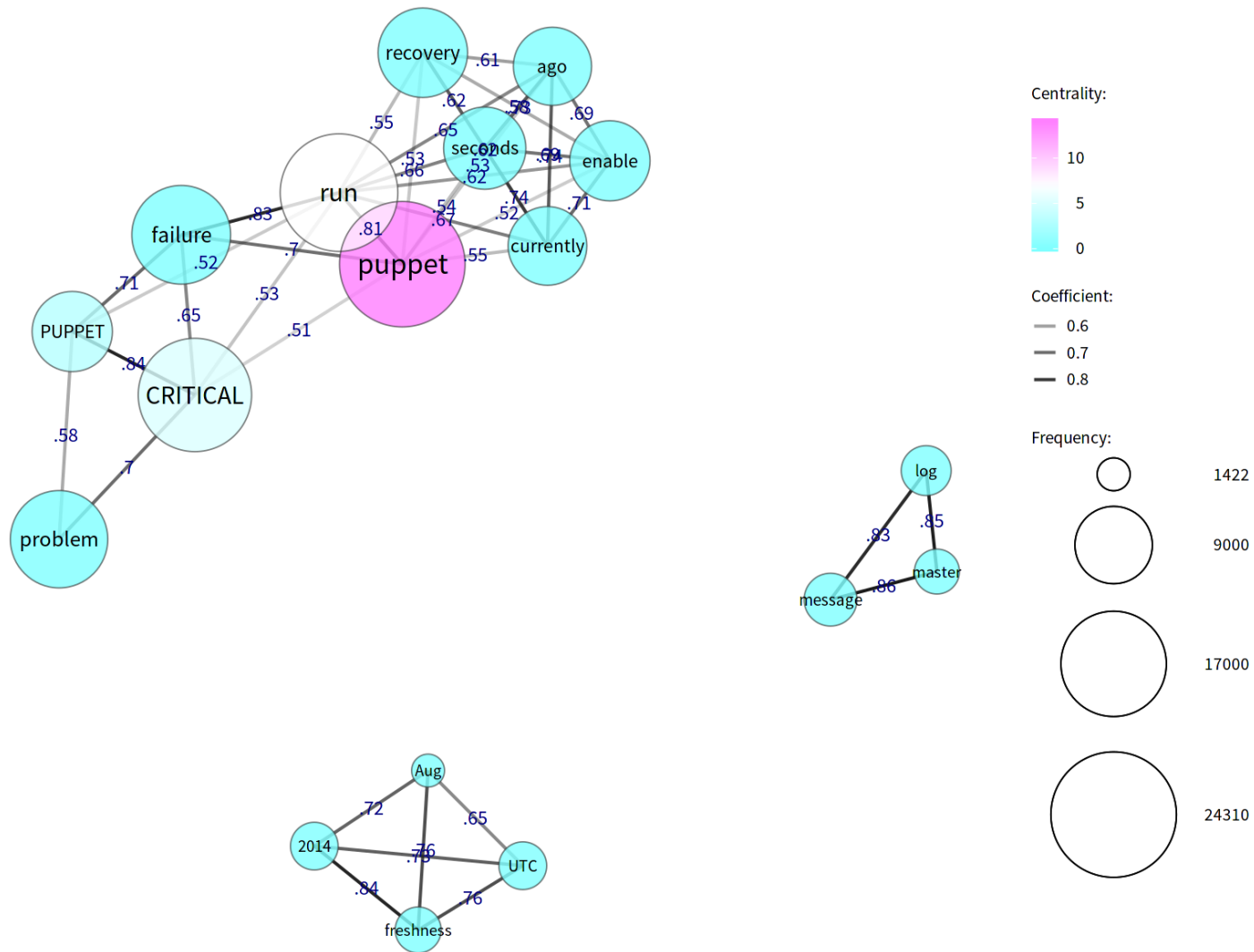
interconnected. In [Figure 8b](#), this pattern is confirmed by the fact that the majority of the most central word-stems are not found among the most frequent, which may give the impression that no one word-stem occupies a conceptually central role – in other words, the configuration of [Figures 8a](#) and [8b](#) is somewhat decentralized. This is by contrast with the channels illustrated in previous sections from [Figure 1a](#) to [Figure 5b](#), in which some elements are at once more frequent and more central.

As for [Figure 6a](#), pertaining to the uniquely configured #ubuntu-offtopic channel, its structure seems to enclose several fairly dissociated central concepts related to eclectic topics, rather than an altogether decentralized structure.

In much the same way as in other diagrams, in [Figure 8a](#), the word-stem “thanks” (in *cluster 02*) and the word-stem “hmm” (in *cluster 11*) are found apart from other elements, at the periphery of the diagram. The same is true of all pseudonyms included except the most frequent, “YuviPanda”<sup>208</sup>. Among other points in common between [Figure 7a](#) and [Figure 8a](#), we also find the word-stems “puppet” and “error” next to each other within the same cluster.

208 Yuvaraj Pandian, an employee of the Wikimedia Foundation. [<http://www.mediawiki.org/wiki/User:Yuvipanda>; [http://en.wikipedia.org/wiki/User:YuviPanda\\_%28WMF%29](http://en.wikipedia.org/wiki/User:YuviPanda_%28WMF%29)]

Figure 8c: “Co-Occurrence Network of Words” diagram representing all publicly logged IRC channels managed by the Wikimedia Foundation except #wikimedia-dev in July and August of 2014, exceptionally including all lines by known “bots” (words = 40; edges = 40).



8. 9 (d) – Interpretation of Figure 8c:

For the sake of consistency, the 15 most frequently occurring word-stems are listed below (in decreasing order of frequency):

“puppet”, “run”, “CRITICAL”, “failure”, “problem”, “owner”, “recovery”, “seconds”, “PUPPET”, “enable”, “currently”, “operations/puppet”, “add”, “log”

Again, the list above is quite different from the most frequent words in other data-sets. Even the few word-stems that are arguably representative of common words in ordinary English conversations – like “run”, “problem”, and “add” – have a specific technical meaning in this context.

Once more, similar to [Figure 7c](#), the participation of “bots” in the channels is reflected in the corresponding “Co-Occurrence Network of Words” diagram by abnormally strong word-associations interconnected in such regular ways that they take the form of fairly symmetric geometrical shapes. Thus, the well-circumscribed clusters in [Figure 8c](#) represent a number of boilerplate expressions, automatically produced and repeated with little variation in the channel.

For example, similar to [Figure 7c](#), we find here the “bot” called “icinga-wm”, responsible for such lines as the ones below:

```
<icinga-wm> PROBLEM - Puppet freshness on osmium is CRITICAL: Last successful Puppet run was Wed 20 Aug 2014 22:49:01 UTC
```

```
<icinga-wm> PROBLEM - Disk space on elastic1005 is CRITICAL: DISK CRITICAL - free space: /var/lib/elasticsearch 20086 MB (3% inode=99%):
```

```
<icinga-wm> PROBLEM - Puppet freshness on db1010 is CRITICAL: Last successful Puppet run was Wed 20 Aug 2014 23:59:40 UTC
```

```
<icinga-wm> RECOVERY - Puppet freshness on db1010 is OK: puppet ran at Thu Aug 21 02:00:10 UTC 2014
```

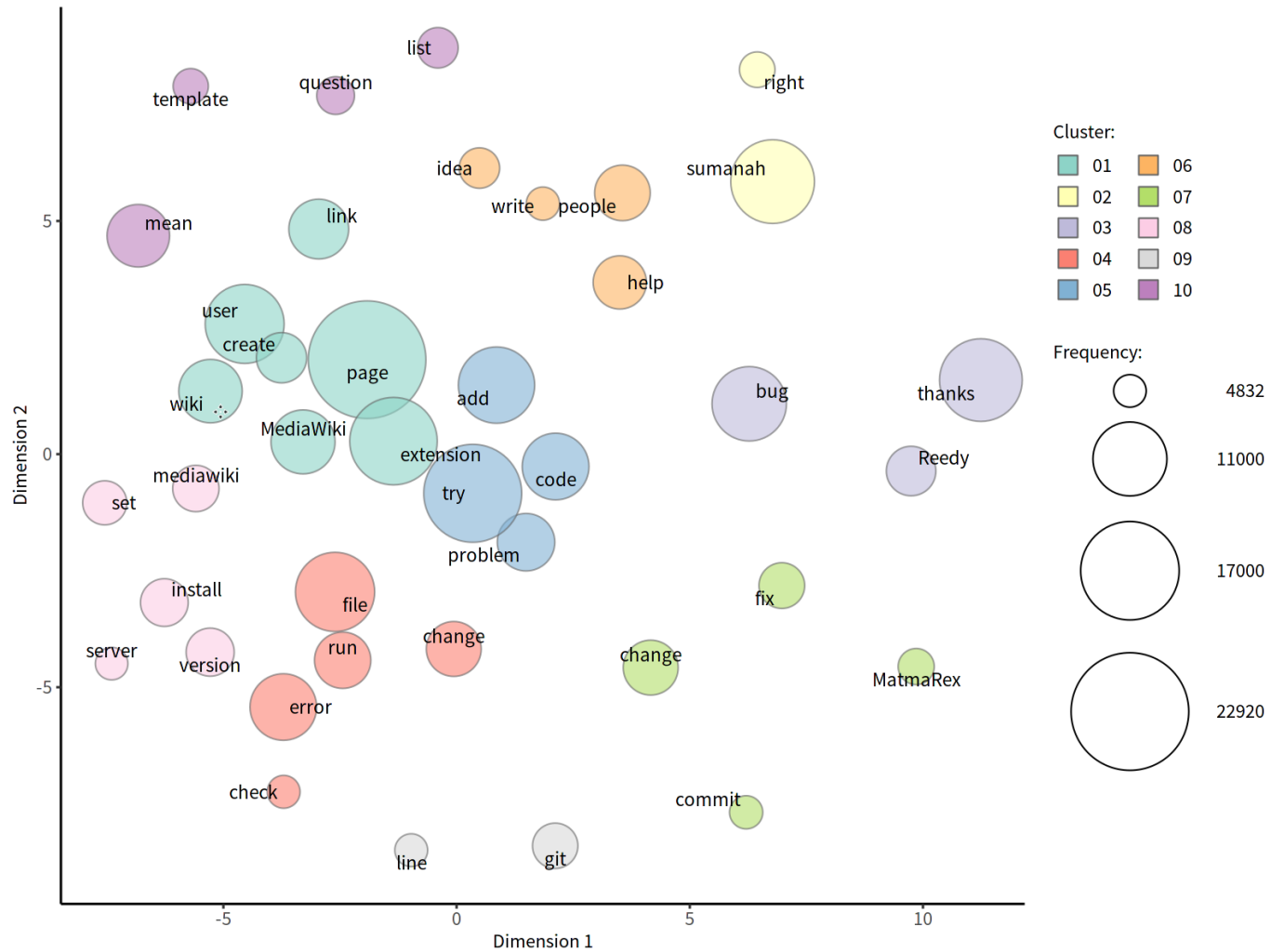
### 8. 9 (e) – Involvement of “bots”:

Figure 7c and Figure 8c are not predominantly maps of human communication, but rather ones of a rather basic variety of “artificial intelligence”: mechanically produced strings of symbols most often make up the dominant part of the conversations. The models nonetheless serve to illustrate an important part of semiosis processes which take place in IRC channels.

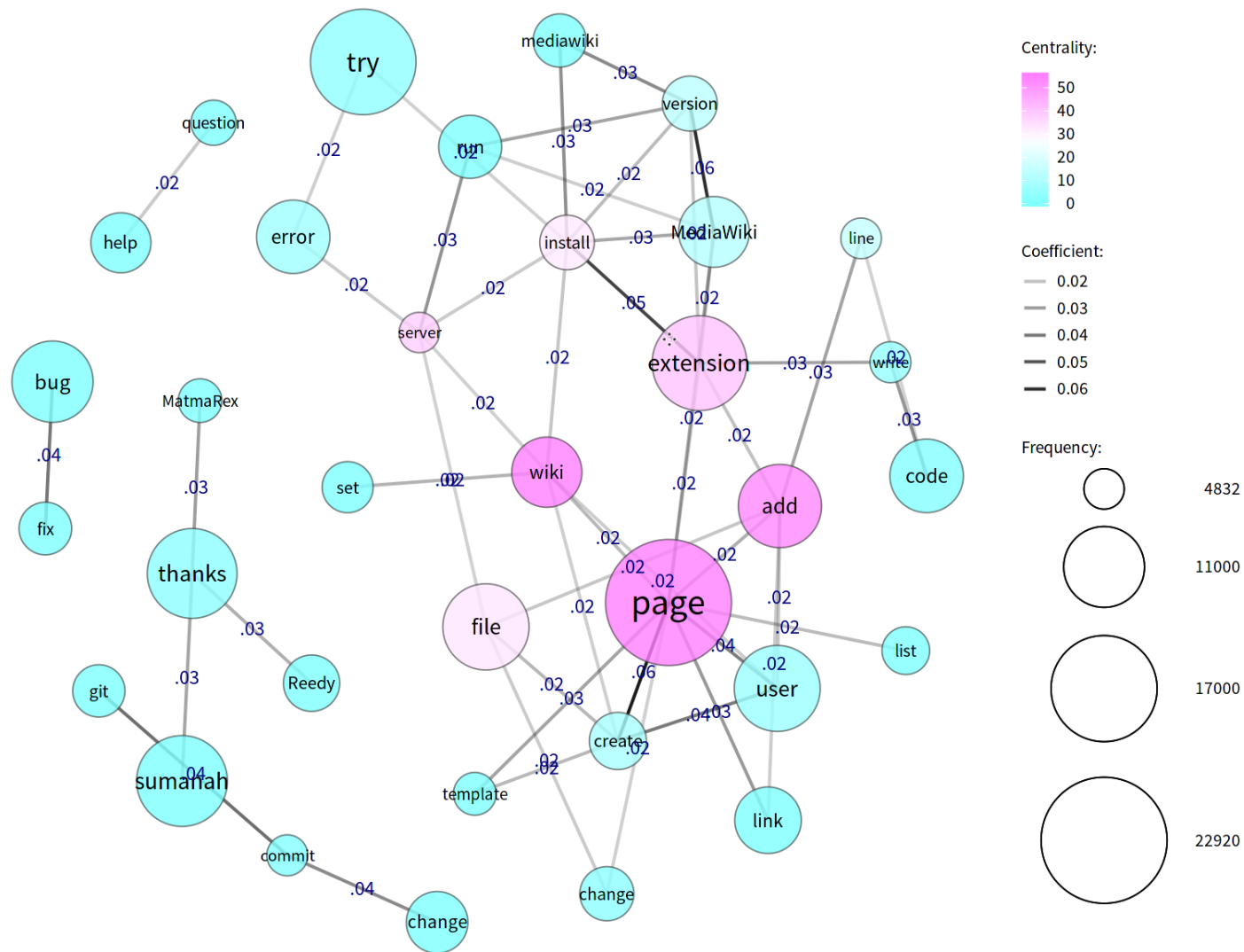
In mechanical word associations, we notice complex symmetries, appearing as unusually regular geometrical shapes, as well as a high level of fragmentation. Such simultaneous regularity, fragmentation, and relatively incongruous choices in sign-associations are characteristic of the form of “artificial intelligence” displayed by most IRC “bots”, and highly un-characteristic of human cognition!

In parallel with his classification of signs, and symmetrically, Peirce distinguished three types of semioses, which Peirce himself did not name but Liska suggested that we call *teleological*, *teleonomic*, and *mechanical* [1996, p. 33]. Only the first corresponds to deliberate intentional use of (conventional) symbols, such as ordinary language use by humans. The second consists in triadic sign-generation of a character defined by the “natural dispositions” of a “quasi-mind” rather than convention and deliberate self-control [CP 4.551 in *Ibid.*], as occurs notably in animal communication. The last – mechanical or *computational* semiosis – corresponds in fact to the *dyadic* action of “quasi-signs” upon one another according to “goals [that] are automatically predetermined or set by agencies external to the system” [Liska, 1996, p. 34], as occurs in the automatic production of a message by a computer program. It is evidently this latter type of semiosis which corresponds to sign-generation by IRC “bots”.

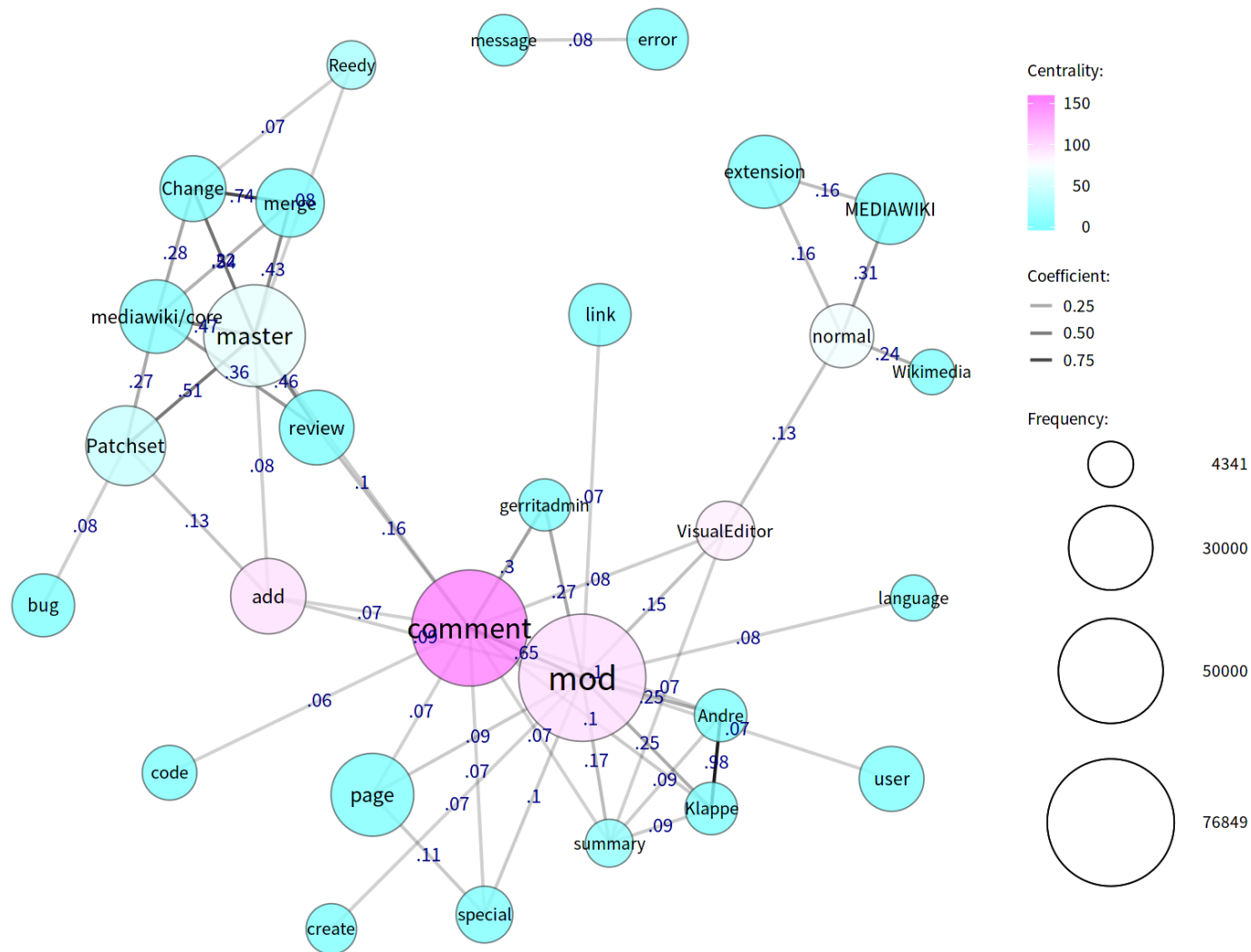
Figure 9a: “Multi-Dimensional Scaling” diagram representing *all publicly logged IRC channels belonging to the MediaWiki project, each of them for the entire period recorded before October 8<sup>th</sup>, 2014* (words = 40, clusters = 10).



**Figure 9b:** “Co-Occurrence Network of Words” diagram representing all publicly logged IRC channels belonging to the MediaWiki project, each of them for the entire period recorded before October 8<sup>th</sup>, 2014 (words = 40; edges = 50).



**Figure 9c:** “Co-Occurrence Network of Words” diagram representing approximately half of the text data from all publicly logged IRC channels belonging to the MediaWiki project, each of them for the entire period recorded before October 8<sup>th</sup>, 2014, exceptionally including all lines by known “bots” (words = 40; edges = 50)



8.9 (f) – Log excerpts corresponding to Figure 9c:

The following excerpt, partly analogous to the strange assemblage of interlinked elements found at the right of Figure 9c, is a rather representative example taken directly from the unprocessed log used to generate the ‘*all channels related to MediaWiki*’ (entire logging period for each channel) data-set:

```
<wikibugs>      (mod) NetworkAuth: Security: Bad IP range recognition -
https://bugzilla.wikimedia.org/38117  summary; +comment (Andre Klapper)

<gerrit-wm>      Change merged: Tobias Gritschacher; [mediawiki/extensions/Wikibase]
(master) - https://gerrit.wikimedia.org/r/27678

<wikibugs>      (mod) CentralAuth account autocreation may be violating WMF's
privacy policy - https://bugzilla.wikimedia.org/40006  +comment (Félix M.
(elfix))

<wikibugs>      (mod) Review Extension:SignupAPI for deployment -
https://bugzilla.wikimedia.org/36225  +comment (Andre Klapper)

<wikibugs>      (NEW) Button is 'pressable' even when disabled -
https://bugzilla.wikimedia.org/40976  normal; VisualEditor: User Interface;
(hartman.wiki)
```

By comparison, the two messages below from the same unprocessed log correspond to most of the cluster at the left of Figure 9c, which roughly consists of six mutually interconnected bubbles:

```
<gerrit-wm>      New patchset: Reedy; "Update Collection to master" [mediawiki/core]
(wmf/1.21wmf1) - https://gerrit.wikimedia.org/r/27654

<gerrit-wm>      Change merged: Reedy; [mediawiki/core] (wmf/1.21wmf1) -
https://gerrit.wikimedia.org/r/27654
```

### **8. 10 – Appendix 10: Additional Diagrams – Including Additional Verbs:**

The exclusion of a broad list of stopwords results in diagrams which tend to comprise more nouns than words of other types. This is useful for the purpose of revealing patterns of sign-use specific to the text data under analysis. However, it is not without severe shortcomings.

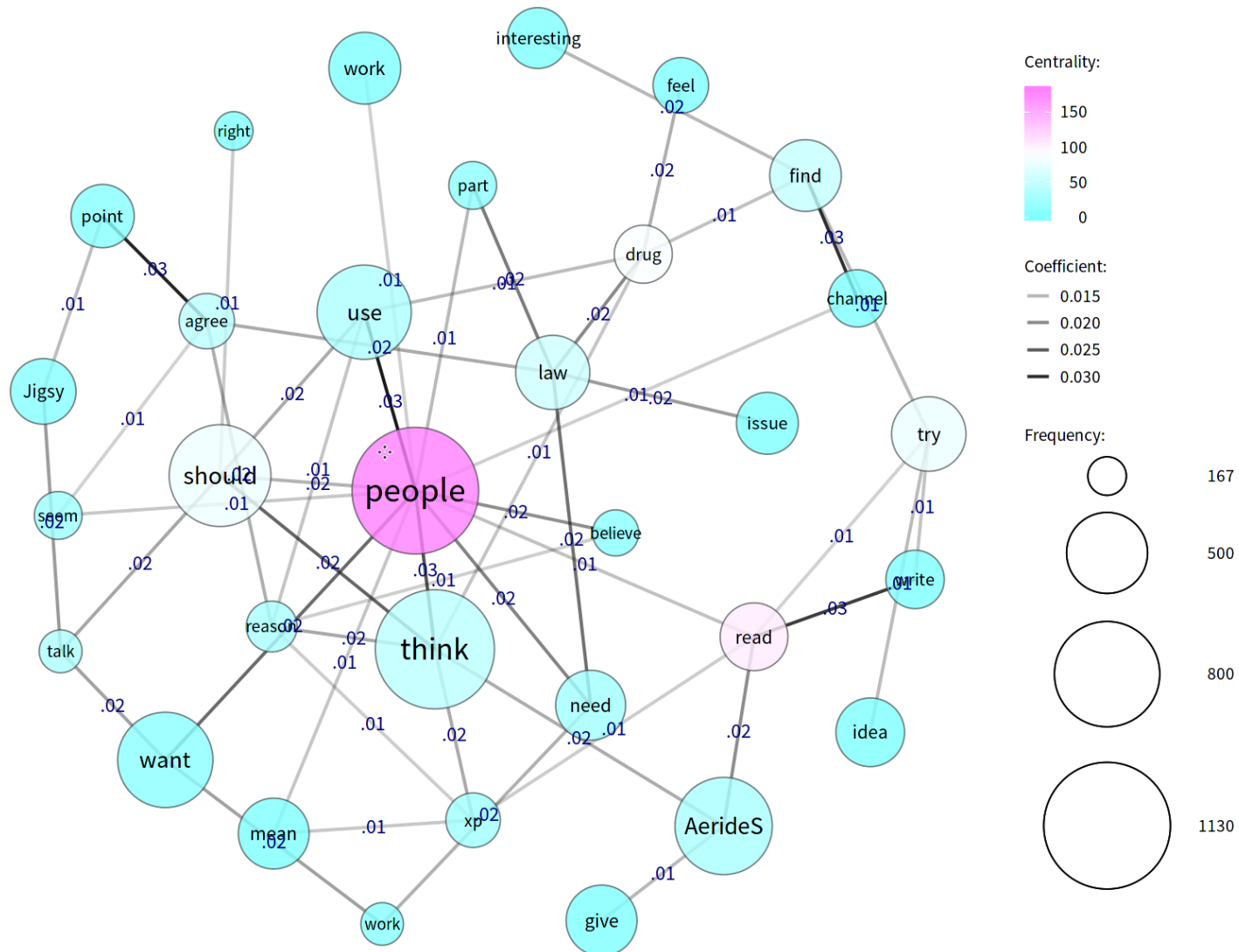
One problem pertains to how, fundamentally, “Co-Occurrence Networks of Words” highlight relationships between pairs of words, and the ways in which these combinations are themselves interconnected. Thus, to elucidate the context in which words – in particular nouns – are used, it was often found useful to include a selection of simple verbs otherwise labelled as stopwords. More generally, it was also found that the inclusion of additional verbs made it easier to shed light on semiotic patterns consistent across many channels.

To illustrate the usefulness of including such verbs in “Co-Occurrence Networks of Words”, and more generally to illustrate the potential pitfalls of stopword selection, this appendix includes a “Co-Occurrence Network of Words” for each one in section 4 – Analysis, as well as for the MediaWiki-related data-set represented in the preceding pages. These diagrams differ from the corresponding ones shown previously in that the only verbs excluded as stopwords are those among the approximately 75 most frequent English words<sup>209</sup>, listed below:

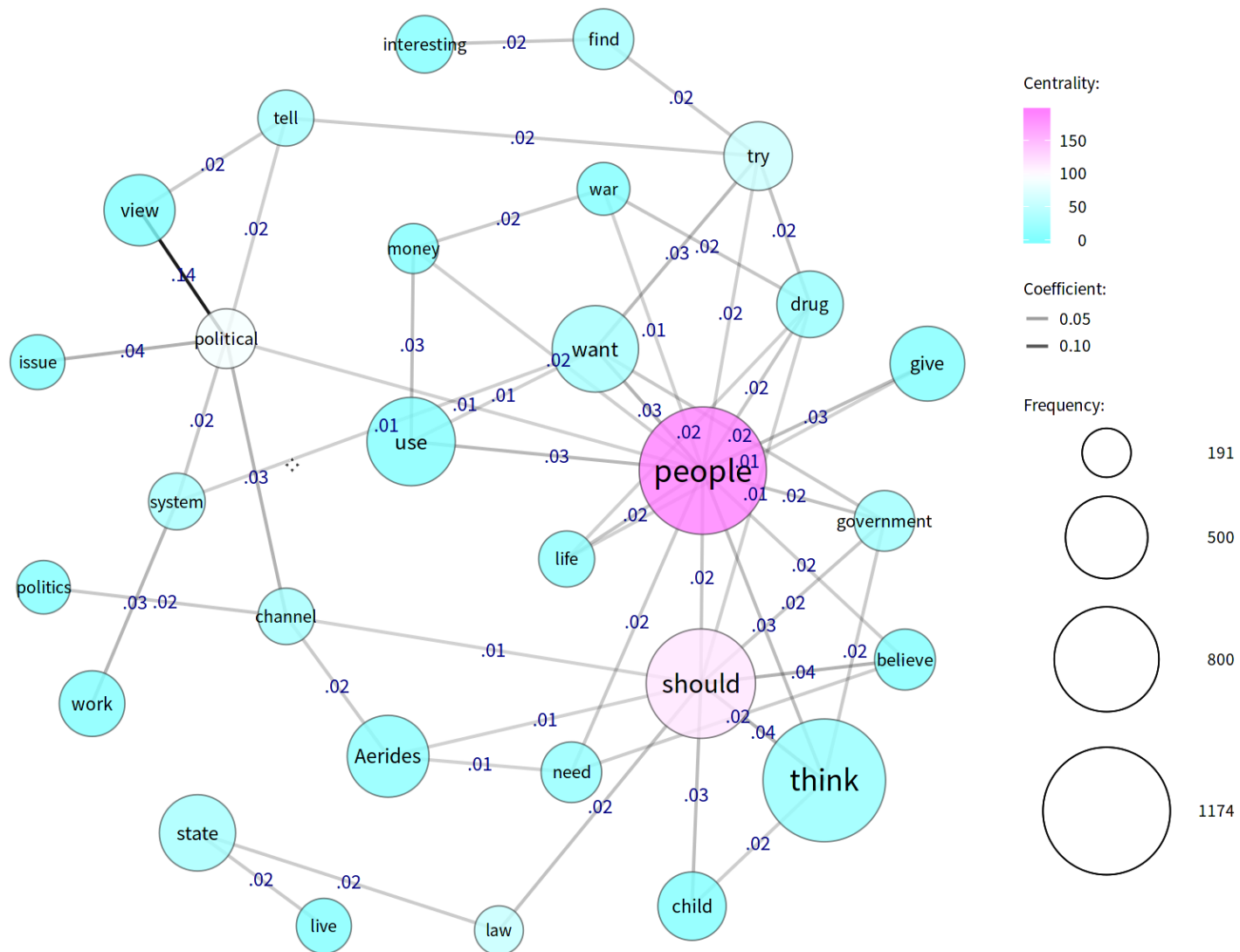
be, have, do, say, will, would, get, go, can, make, know, take, could, see, come

209 The Wikipedia page [https://en.wikipedia.org/wiki/Most\\_common\\_words\\_in\\_English](https://en.wikipedia.org/wiki/Most_common_words_in_English) includes two lists – one based on the 2+ billion word “Oxford English Corpus”, and another based on the 450+ million word “Corpus of Contemporary American English” (which is calculated according to word dispersion as well as word frequency). Here, an average of the word's rank in each of the lists was used.

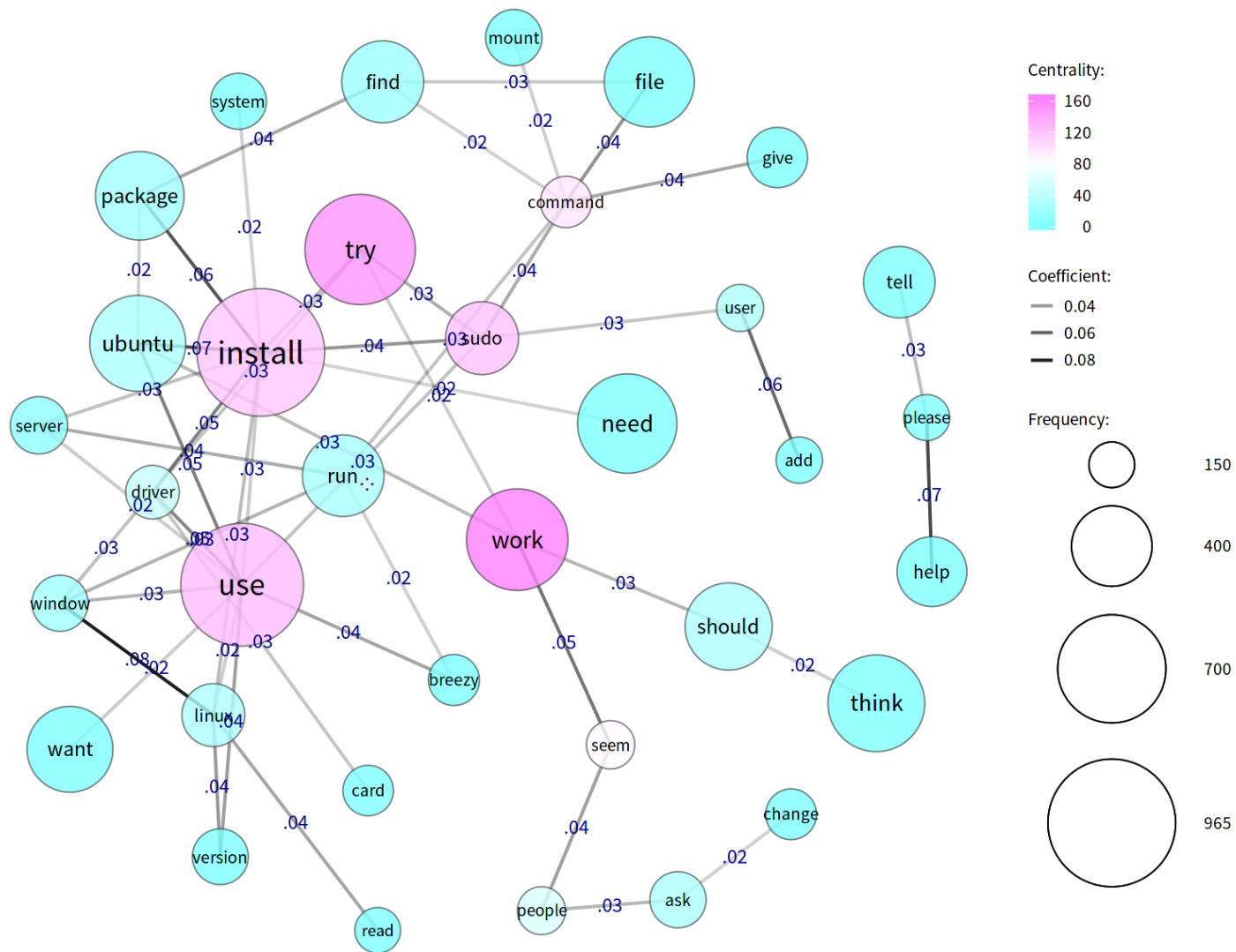
**Figure I:** “Co-Occurrence Network of Words” diagram representing the #libertyinmind IRC channel from October 29th, 2013 to August 4th, 2014, including additional verbs. (words = 40; edges = 50)



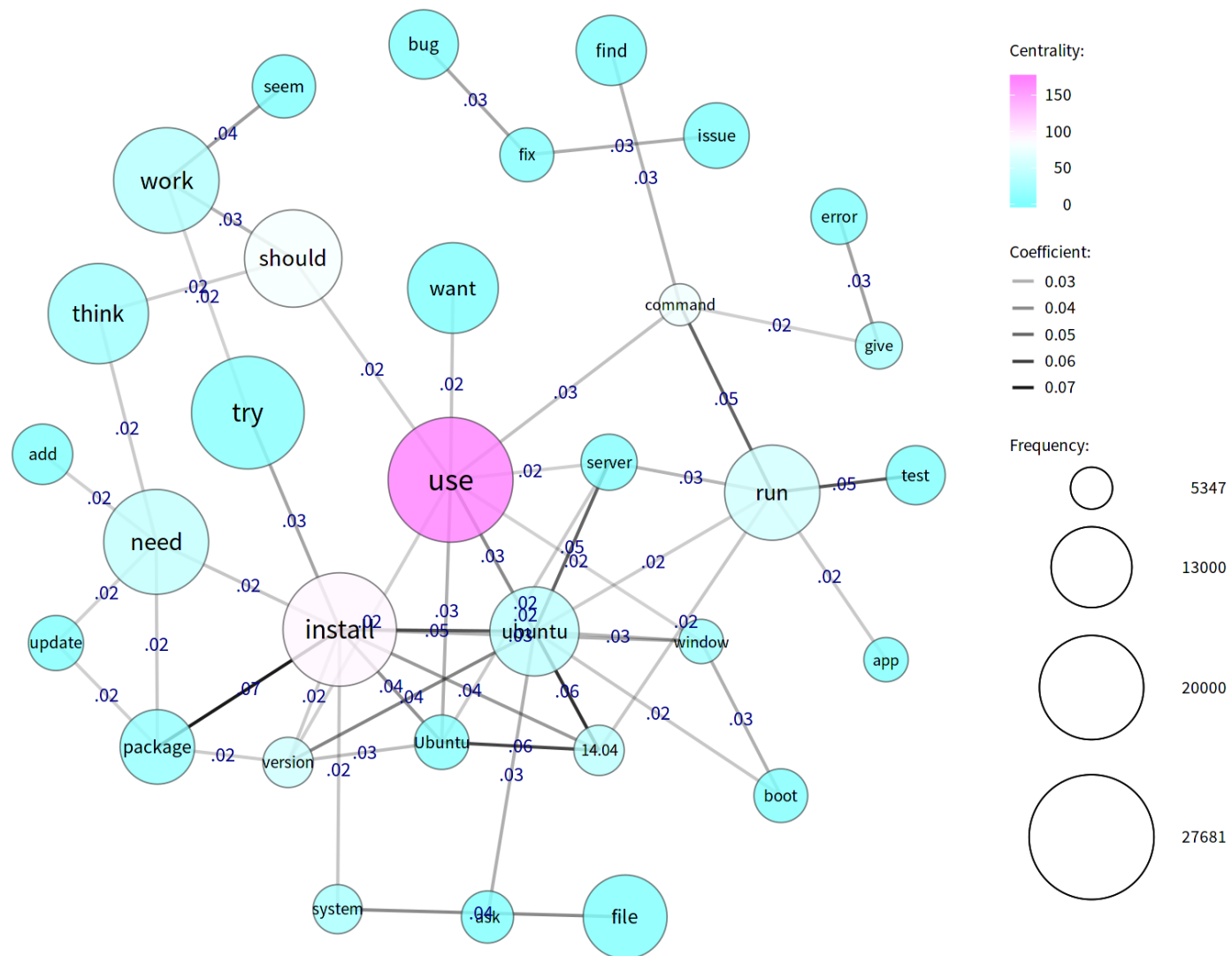
**Figure II:** “Co-Occurrence Network of Words” diagram representing the #politics IRC channel from April 29th, 2012 to August 4th, 2014, including additional verbs. (words = 40; edges = 50)



**Figure III:** “Co-Occurrence Network of Words” diagram representing all IRC channels belonging to the Ubuntu project that were publicly logged on December 25th, 2004, including additional verbs. (words = 41; edges = 50)

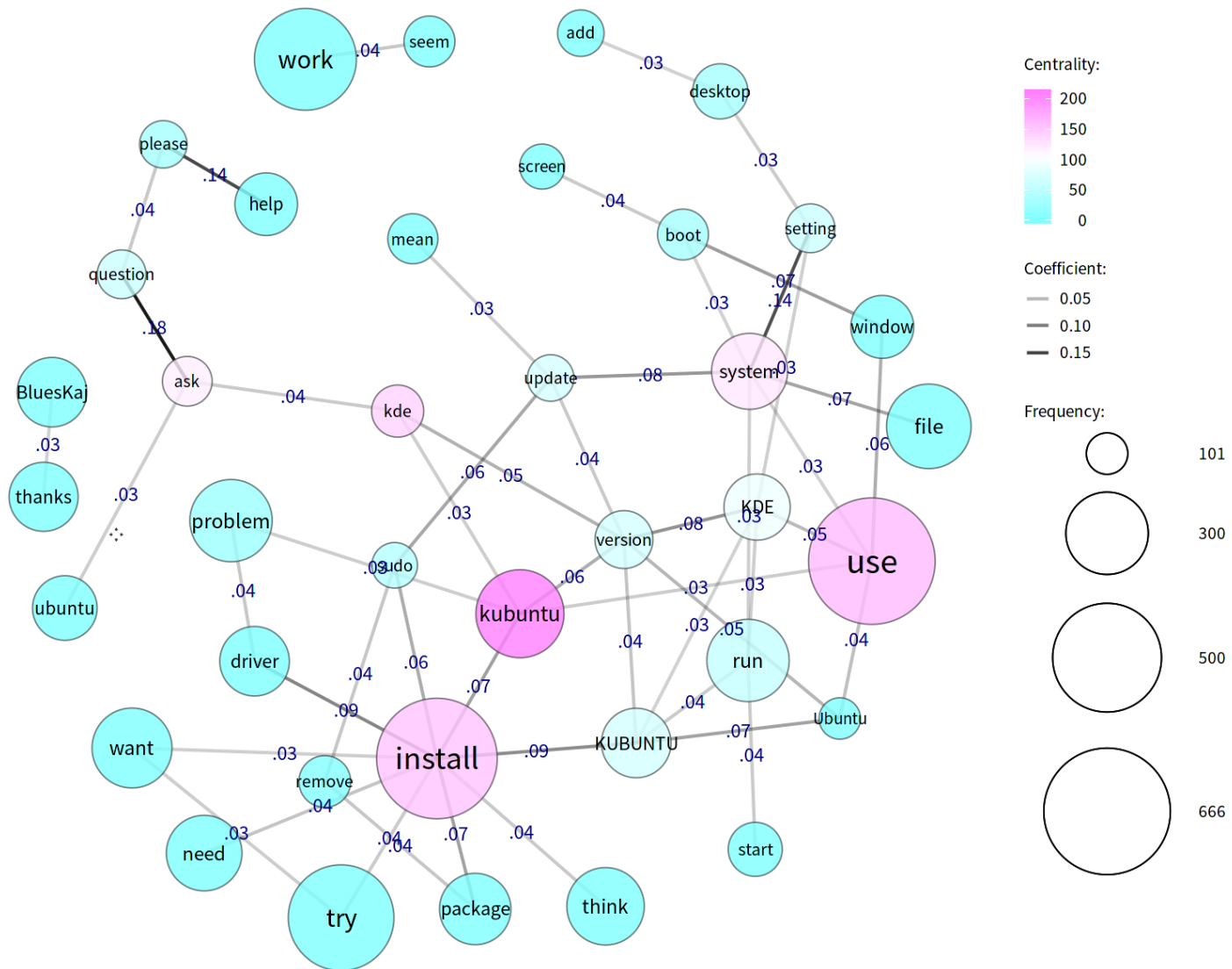


**Figure IV:** “Co-Occurrence Network of Words” diagram representing all English-language IRC channels belonging to the Ubuntu project that were publicly logged in July and August of 2014, including additional verbs. (words = 40; edges = 50)

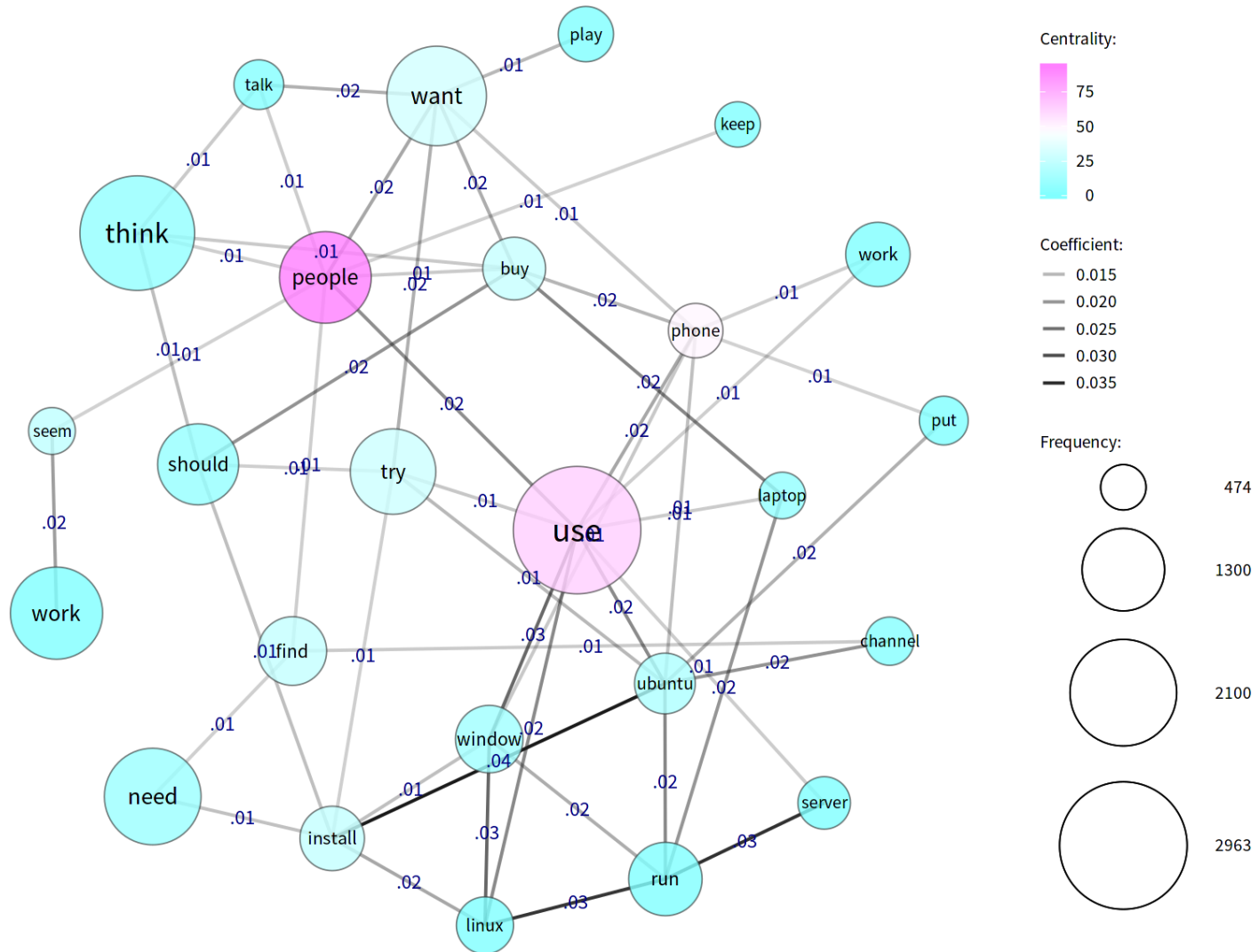


**Figure V:** “Co-Occurrence Network of Words” diagram representing the #kubuntu IRC channel in January of 2013, including additional verbs.

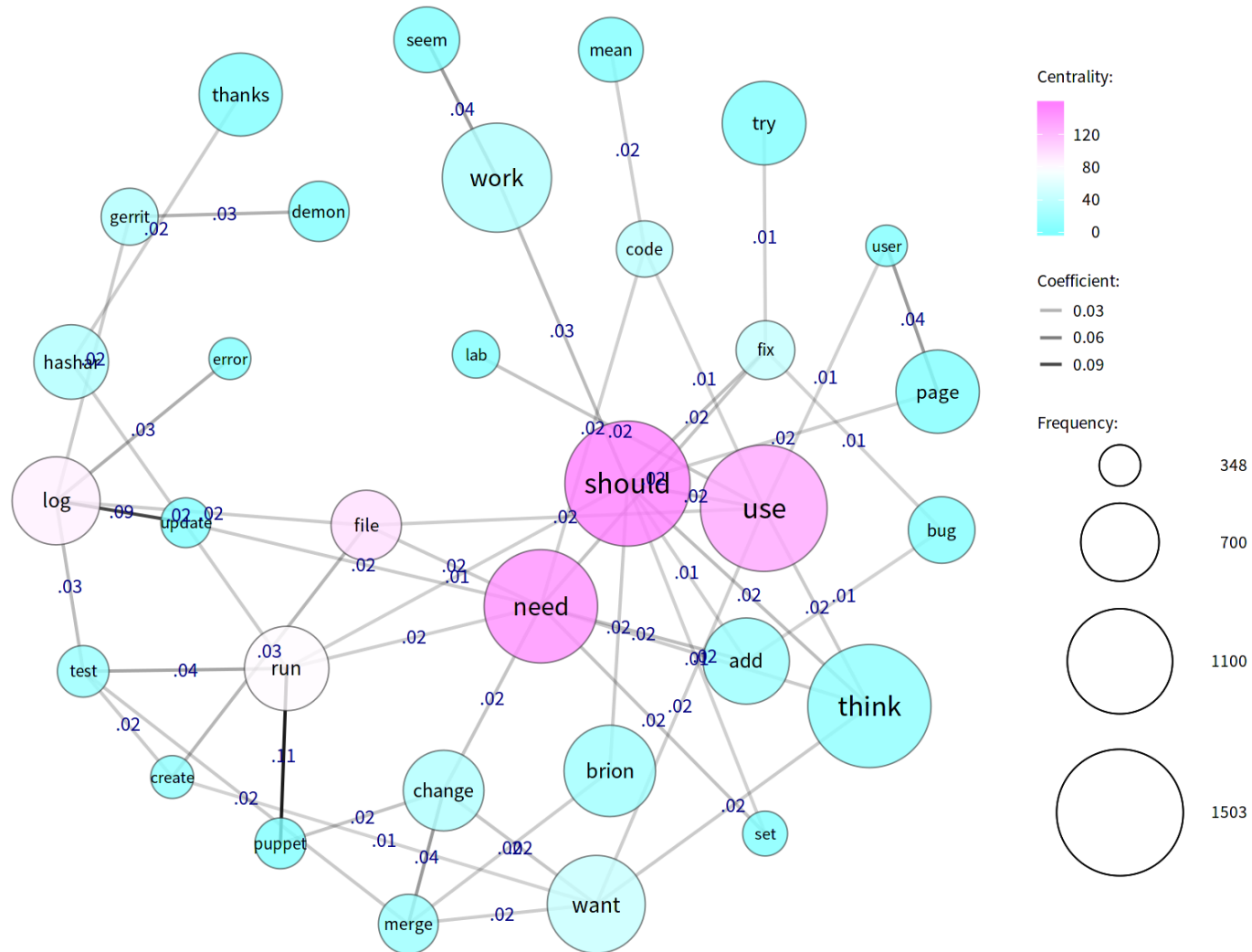
(words = 41; edges = 50)



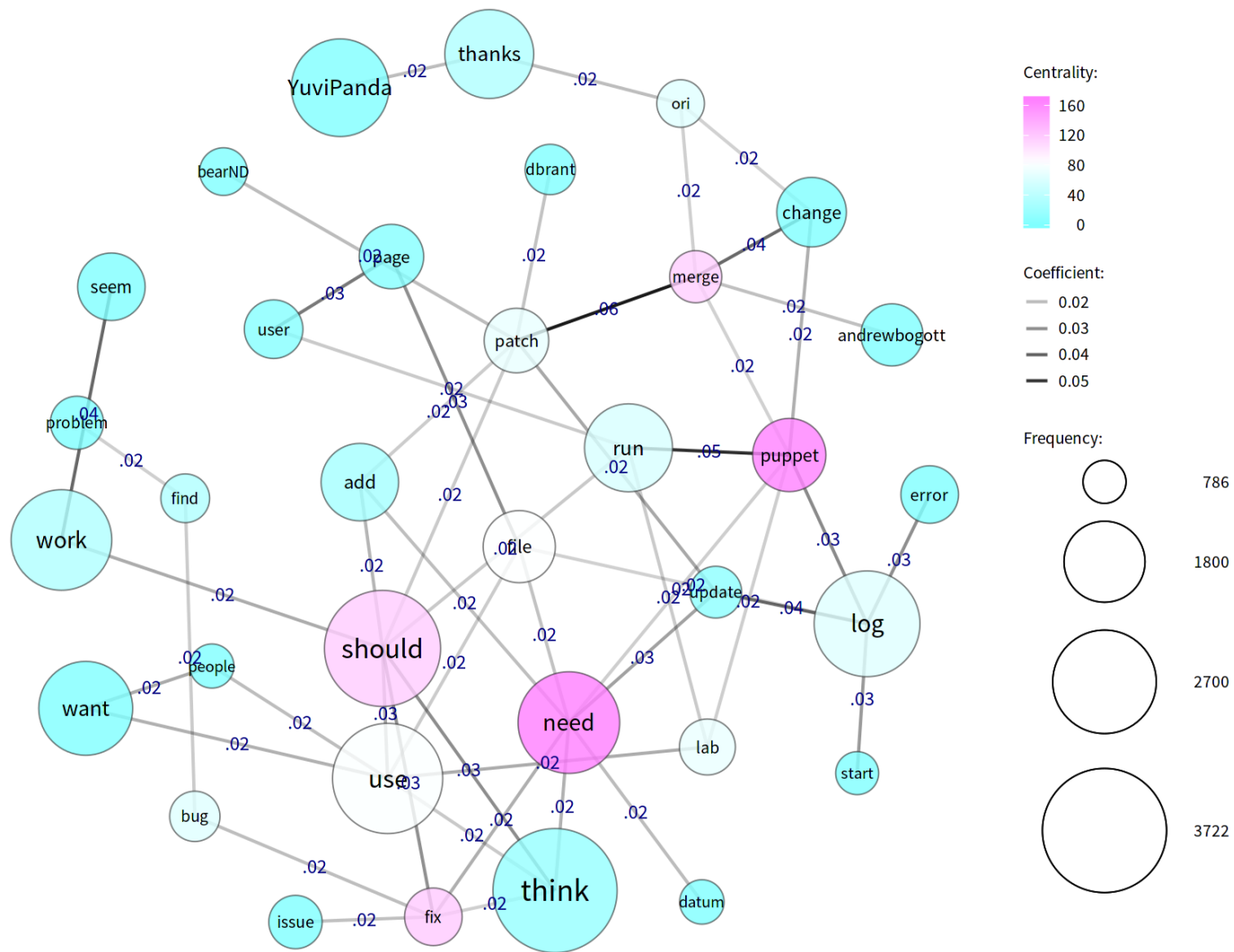
**Figure VI:** “Co-Occurrence Network of Words” diagram representing the #ubuntu-offtopic IRC channel from August 8th, 2014 to October 13th, 2014, including additional verbs. (words = 40; edges = 50)



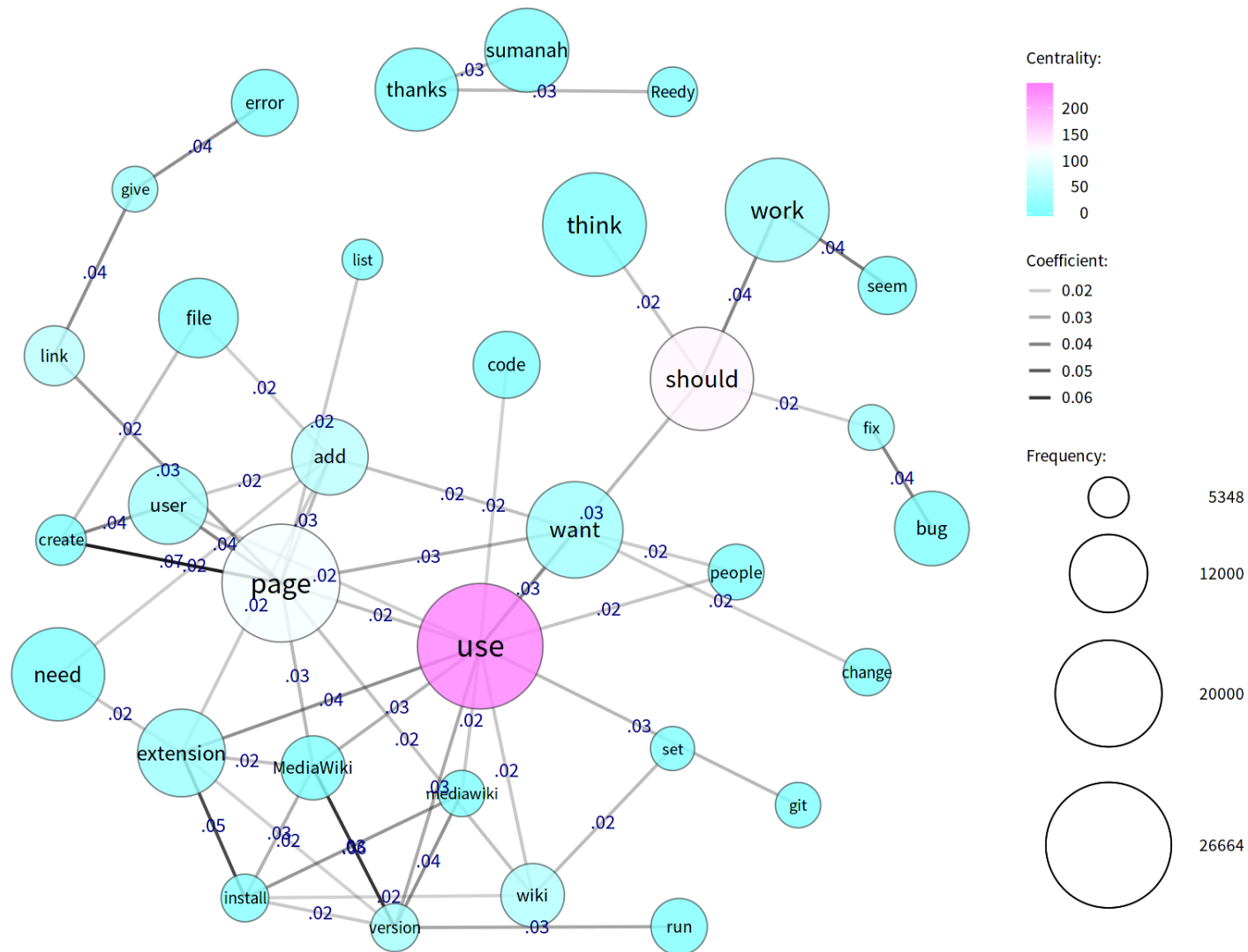
**Figure VII:** “Co-Occurrence Network of Words” diagram representing the #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office and #wikimedia-operations IRC channels in June of 2013, including additional verbs. (words = 40; edges = 50)



**Figure VIII:** “Co-Occurrence Network of Words” diagram representing all publicly logged IRC channels managed by the Wikimedia Foundation except #wikimedia-dev in July and August of 2014, including additional verbs. (words = 40; edges = 50)



**Figure IX:** “Co-Occurrence Network of Words” diagram representing all publicly logged IRC channels belonging to the MediaWiki project, each of them for the entire period recorded before October 8<sup>th</sup>, 2014, including additional verbs. (words = 40; edges = 50)



### **8. 11 – Appendix 11: General Features of the Data Analyzed**

In all cases, the information listed corresponds to the text data after the removal of time-stamps and that of speakers' pseudonyms preceding each message, as well as non-standard characters and lines containing too many such characters. When appropriate, and unless specified, it also corresponds to the data after deletion of all lines by known “bots” (more specifically, this applies to the data corresponding to Figures [4a](#), [4b](#), [7a](#), [7b](#), [8a](#), [8b](#), [9a](#), [9b](#), [10](#), and [11](#)).

| <b>Figures 1a, 1b &amp; I</b> |                                        |
|-------------------------------|----------------------------------------|
| <i>Channel(s) / network:</i>  | #libertyinmind / Freenode              |
| <i>Period:</i>                | 2013-10-29 to 2014-08-04 (my own logs) |
| <i>Size:</i>                  | ≈ 1.6 MiB (1,676,540 bytes)            |
| <i>Lines:</i>                 | 30,367                                 |
| <i>Words:</i>                 | 271,444                                |
| <i>Characters:</i>            | 1,645,837                              |

| <b>Figures 2a, 2b &amp; II</b> |                                        |
|--------------------------------|----------------------------------------|
| <i>Channel(s) / network:</i>   | #politics / Rizon                      |
| <i>Period:</i>                 | 2012-04-29 to 2014-08-04 (my own logs) |
| <i>Size:</i>                   | ≈ 1.5 MiB (1,567,441 bytes)            |
| <i>Lines:</i>                  | 31,863                                 |
| <i>Words:</i>                  | 264,456                                |
| <i>Characters:</i>             | 1,535,177                              |

| <b>Figures 3a, 3b &amp; III</b> |                                                                                           |
|---------------------------------|-------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i>    | All channels related to Ubuntu listed in <a href="#">6. 2 – Appendix 2 [1]</a> / Freenode |
| <i>Period:</i>                  | 2004-12-25 (public logs)                                                                  |
| <i>Size:</i>                    | ≈ 0.9 MiB (928,075 bytes)                                                                 |
| <i>Lines:</i>                   | 18,975                                                                                    |

|                    |         |
|--------------------|---------|
| <i>Words:</i>      | 158,060 |
| <i>Characters:</i> | 909,100 |

| <b>Figures 4a, 4b &amp; IV</b> |                                                                                                                                 |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i>   | All channels related to Ubuntu listed in <a href="#">6. 2 – Appendix 2 [1]</a> (only English-language conversations) / Freenode |
| <i>Period:</i>                 | 2014-07-01 to 2014-08-31 (public logs)                                                                                          |
| <i>Size:</i>                   | ≈ 32.3 MiB (33,869,286 bytes)                                                                                                   |
| <i>Lines:</i>                  | 582,416                                                                                                                         |
| <i>Words:</i>                  | 5,800,369                                                                                                                       |
| <i>Characters:</i>             | 33,286,870                                                                                                                      |

| <b>Figures 5a, 5b &amp; V</b> |                                        |
|-------------------------------|----------------------------------------|
| <i>Channel(s) / network:</i>  | #kubuntu / Freenode                    |
| <i>Period:</i>                | 2013-01-01 to 2013-01-31 (public logs) |
| <i>Size:</i>                  | ≈ 0.52 MiB (534,110)                   |
| <i>Lines:</i>                 | 9,691                                  |
| <i>Words:</i>                 | 93,266                                 |
| <i>Characters:</i>            | 524,419                                |

| <b>Figures 6a, 6b &amp; VI</b> |                                        |
|--------------------------------|----------------------------------------|
| <i>Channel(s) / network:</i>   | #ubuntu-offtopic / Freenode            |
| <i>Period:</i>                 | 2014-08-08 to 2014-10-13 (my own logs) |
| <i>Size:</i>                   | ≈ 3.7 MiB (3,922,549 bytes)            |
| <i>Lines:</i>                  | 99,078                                 |
| <i>Words:</i>                  | 705,598                                |
| <i>Characters:</i>             | 3,823,471                              |

| <b>Figures 7a, 7b &amp; VII</b> |                                                                                                                                                       |
|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i>    | #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office, #wikimedia-operations / Freenode |
| <i>Period:</i>                  | 2013-06-01 to 2013-06-30 (public logs)                                                                                                                |
| <i>Size:</i>                    | ≈ 2.2 MiB (2,277,592 bytes)                                                                                                                           |
| <i>Lines:</i>                   | 45,684                                                                                                                                                |
| <i>Words:</i>                   | 373,316                                                                                                                                               |
| <i>Characters:</i>              | 2,231,908                                                                                                                                             |

| <b>Figure 7c</b>             |                                                                                                                                                                                             |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i> | #wikimedia-analytics, #wikimedia-dev, #wikimedia-fundraising, #wikimedia-labs, #wikimedia-mobile, #wikimedia-office, #wikimedia-operations / Freenode<br>(Including lines by known “bots”.) |
| <i>Period:</i>               | 2013-06-01 to 2013-06-30 (public logs)                                                                                                                                                      |
| <i>Size:</i>                 | ≈ 5.2 MiB (5,415,211 bytes)                                                                                                                                                                 |
| <i>Lines:</i>                | 73,780                                                                                                                                                                                      |
| <i>Words:</i>                | 782,714                                                                                                                                                                                     |
| <i>Characters:</i>           | 5,341,431                                                                                                                                                                                   |

| <b>Figures 8a, 8b &amp; VIII</b> |                                                                                                                    |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i>     | All channels related to Wikimedia listed in <a href="#">6. 2 – Appendix 2 [2]</a> except #wikimedia-dev / Freenode |
| <i>Period:</i>                   | 2014-07-01 to 2014-08-31; 2014-08-01 to 2014-08-31 for #wikimedia-research (public logs)                           |
| <i>Size:</i>                     | ≈ 5.0 MiB (5,233,824 bytes)                                                                                        |
| <i>Lines:</i>                    | 95,961                                                                                                             |
| <i>Words:</i>                    | 866,431                                                                                                            |
| <i>Characters:</i>               | 5,137,863                                                                                                          |

| Figure 8c                    |                                                                                                                                                          |
|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i> | All channels related to Wikimedia listed in <u>6. 2 – Appendix 2 [2]</u> except #wikimedia-dev / Freenode<br>( <u>Including lines by known “bots”.</u> ) |
| <i>Period:</i>               | 2014-07-01 to 2014-08-31; 2014-08-01 to 2014-08-31 for #wikimedia-research (public logs)                                                                 |
| <i>Size:</i>                 | ≈ 10.5 MiB (10,964,607 bytes)                                                                                                                            |
| <i>Lines:</i>                | 137,288                                                                                                                                                  |
| <i>Words:</i>                | 1,573,424                                                                                                                                                |
| <i>Characters:</i>           | 10,827,319                                                                                                                                               |

| Figures 9a, 9b & IX          |                                                                                                        |
|------------------------------|--------------------------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i> | All channels related to MediaWiki listed in <u>6. 2 – Appendix 2 [2]</u> / Freenode                    |
| <i>Period:</i>               | Various: for the channels publicly logged for the longest, from 2010-08-25 to 2014-10-07 (public logs) |
| <i>Size:</i>                 | ≈ 30.4 MiB (31,911,024 bytes)                                                                          |
| <i>Lines:</i>                | 548,489                                                                                                |
| <i>Words:</i>                | 5,251,548                                                                                              |
| <i>Characters:</i>           | 31,362,535                                                                                             |

| Figure 9c                    |                                                                                                                                    |                                                                                       |
|------------------------------|------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| <i>Channel(s) / network:</i> | All channels related to MediaWiki listed in <u>6. 2 – Appendix 2 [2]</u> / Freenode<br>( <u>Including lines by known “bots”.</u> ) | <u>Due to computer hardware limitations (insufficient RAM), the resulting log was</u> |
| <i>Period:</i>               | Various: for the channels publicly logged for the longest, from 2010-08-25 to 2014-10-07 (public logs)                             | <u>cut by approximately half between randomly selected lines.</u>                     |
| <i>Size:</i>                 | ≈ 30.0 MiB (31,405,458 bytes)                                                                                                      |                                                                                       |
| <i>Lines:</i>                | 350,229                                                                                                                            |                                                                                       |
| <i>Words:</i>                | 3,955,054                                                                                                                          |                                                                                       |
| <i>Characters:</i>           | 30,710,784                                                                                                                         |                                                                                       |

**8. 12 – Appendix 12: Sample Table of Word Frequencies**

*Sample frequency list for the 40 most common word-stems in the #politics channel from April 29th of 2012 to August 4th of 2014. This data corresponds to Figures 2a, 2b, and II.*

| Words       | Frequency |
|-------------|-----------|
| people      | 1174      |
| lol         | 588       |
| Aerides     | 405       |
| himarichan  | 364       |
| state       | 355       |
| mean        | 335       |
| view        | 314       |
| try         | 296       |
| agree       | 295       |
| child       | 294       |
| drug        | 282       |
| believe     | 248       |
| government  | 246       |
| political   | 241       |
| point       | 231       |
| interesting | 229       |
| system      | 225       |
| channel     | 224       |
| life        | 222       |
| tell        | 222       |
| live        | 218       |
| issue       | 216       |
| politics    | 210       |
| war         | 207       |
| money       | 195       |
| idea        | 192       |
| lpf         | 192       |
| part        | 192       |
| happen      | 191       |
| law         | 191       |
| read        | 190       |
| speech      | 188       |

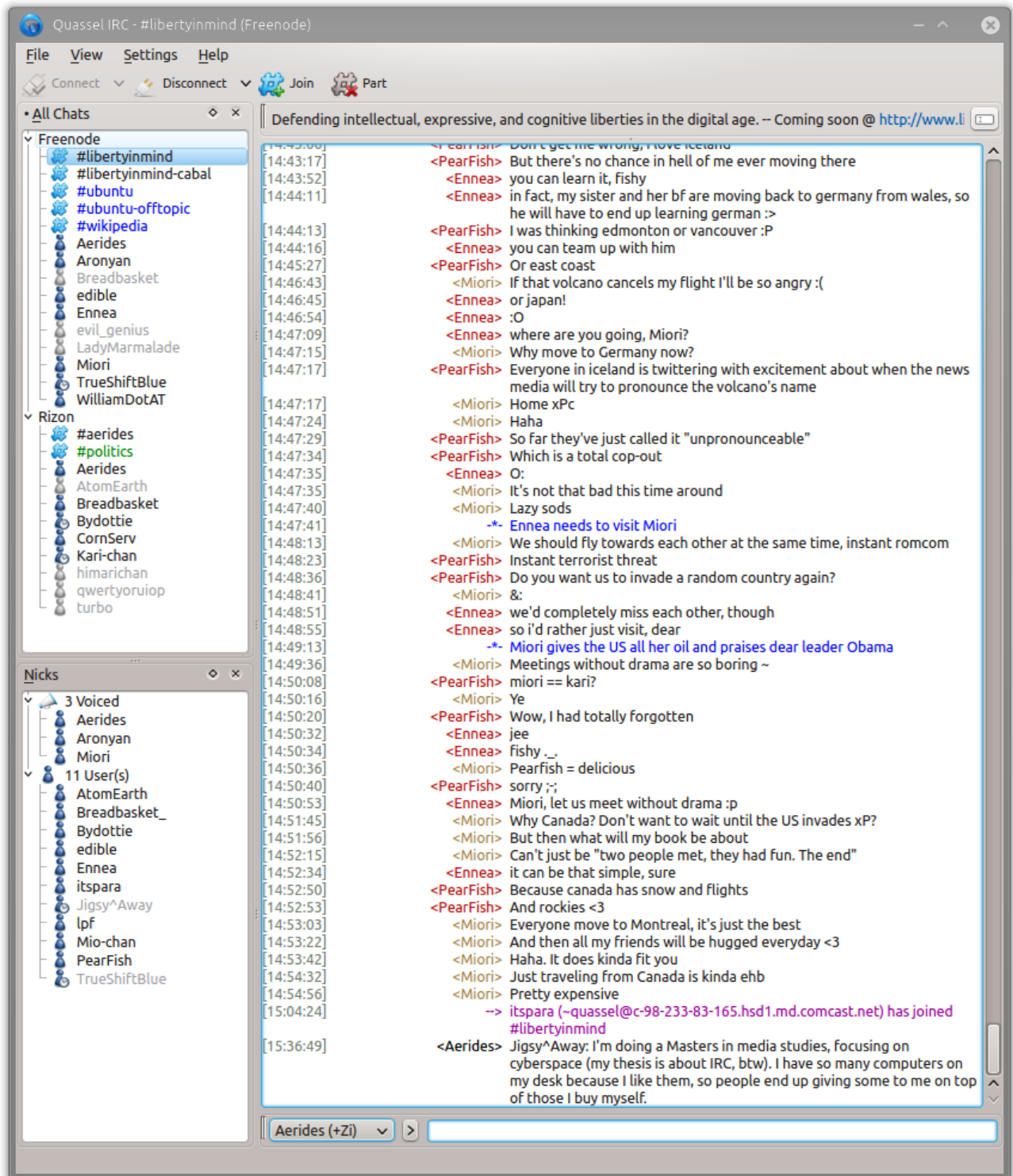
### 8.13 – Appendix 13: Sample of Similarity Matrix (Jaccard Similarity Coefficient)

Sample “Similarity Matrix” for the 28 most common word-stems in the #politics channel from April 29th of 2012 to August 4th of 2014.

This includes only part of the data shown in the corresponding visualizations ([Figure 2a](#), [Figure 2b](#), and [Figure II](#)), which used the 40 most common word-stems. Obviously, data of such complexity is excruciatingly difficult to interpret without a diagrammatic summary.

|              | *people | *lol  | *Aerides | *himarichan | *state | *mean | *vew  | *try  | *agree | *child | *drug | *believe | *government | *political | *point | *interesting | *system | *channel | *life | *tell | *live | *issue | *politics | *war  | *money | *idea | *lpf  | *part |
|--------------|---------|-------|----------|-------------|--------|-------|-------|-------|--------|--------|-------|----------|-------------|------------|--------|--------------|---------|----------|-------|-------|-------|--------|-----------|-------|--------|-------|-------|-------|
| *people      | 1.000   | 0.005 | 0.010    | 0.008       | 0.015  | 0.016 | 0.020 | 0.017 | 0.008  | 0.017  | 0.039 | 0.016    | 0.023       | 0.016      | 0.015  | 0.009        | 0.015   | 0.017    | 0.029 | 0.010 | 0.015 | 0.012  | 0.011     | 0.015 | 0.020  | 0.015 | 0.002 | 0.013 |
| *lol         | 0.005   | 1.000 | 0.001    | 0.009       | 0.002  | 0.004 | 0.000 | 0.007 | 0.002  | 0.001  | 0.003 | 0.000    | 0.001       | 0.000      | 0.001  | 0.001        | 0.000   | 0.001    | 0.000 | 0.001 | 0.000 | 0.001  | 0.005     | 0.001 | 0.001  | 0.004 | 0.006 | 0.000 |
| *Aerides     | 0.010   | 0.001 | 1.000    | 0.004       | 0.006  | 0.011 | 0.042 | 0.006 | 0.006  | 0.017  | 0.034 | 0.016    | 0.005       | 0.017      | 0.008  | 0.006        | 0.005   | 0.024    | 0.003 | 0.016 | 0.002 | 0.005  | 0.010     | 0.002 | 0.008  | 0.008 | 0.008 | 0.005 |
| *himarichan  | 0.008   | 0.009 | 0.004    | 1.000       | 0.008  | 0.004 | 0.010 | 0.006 | 0.012  | 0.005  | 0.006 | 0.005    | 0.002       | 0.002      | 0.005  | 0.013        | 0.002   | 0.006    | 0.002 | 0.003 | 0.003 | 0.005  | 0.005     | 0.008 | 0.002  | 0.007 | 0.003 | 0.003 |
| *state       | 0.015   | 0.002 | 0.006    | 0.008       | 1.000  | 0.012 | 0.000 | 0.005 | 0.003  | 0.015  | 0.003 | 0.012    | 0.008       | 0.000      | 0.005  | 0.002        | 0.009   | 0.000    | 0.009 | 0.003 | 0.017 | 0.009  | 0.005     | 0.009 | 0.024  | 0.007 | 0.004 | 0.007 |
| *mean        | 0.016   | 0.004 | 0.011    | 0.004       | 0.012  | 1.000 | 0.006 | 0.006 | 0.010  | 0.014  | 0.008 | 0.009    | 0.009       | 0.004      | 0.003  | 0.004        | 0.007   | 0.007    | 0.005 | 0.002 | 0.004 | 0.007  | 0.000     | 0.007 | 0.006  | 0.004 | 0.006 | 0.006 |
| *vew         | 0.020   | 0.000 | 0.042    | 0.010       | 0.000  | 0.006 | 1.000 | 0.013 | 0.008  | 0.018  | 0.055 | 0.004    | 0.013       | 0.115      | 0.026  | 0.013        | 0.002   | 0.035    | 0.010 | 0.045 | 0.006 | 0.030  | 0.018     | 0.004 | 0.002  | 0.012 | 0.006 | 0.006 |
| *try         | 0.017   | 0.007 | 0.006    | 0.006       | 0.005  | 0.006 | 0.013 | 1.000 | 0.007  | 0.000  | 0.022 | 0.000    | 0.006       | 0.013      | 0.017  | 0.004        | 0.004   | 0.004    | 0.006 | 0.018 | 0.006 | 0.002  | 0.006     | 0.000 | 0.002  | 0.008 | 0.008 | 0.004 |
| *agree       | 0.008   | 0.002 | 0.006    | 0.012       | 0.003  | 0.010 | 0.008 | 0.007 | 1.000  | 0.015  | 0.004 | 0.010    | 0.009       | 0.018      | 0.011  | 0.012        | 0.002   | 0.004    | 0.004 | 0.002 | 0.000 | 0.012  | 0.000     | 0.006 | 0.004  | 0.011 | 0.004 | 0.013 |
| *child       | 0.017   | 0.001 | 0.017    | 0.005       | 0.015  | 0.014 | 0.018 | 0.000 | 0.015  | 1.000  | 0.022 | 0.032    | 0.002       | 0.004      | 0.010  | 0.008        | 0.027   | 0.000    | 0.022 | 0.004 | 0.013 | 0.020  | 0.000     | 0.002 | 0.019  | 0.012 | 0.009 | 0.007 |
| *drug        | 0.039   | 0.003 | 0.034    | 0.006       | 0.003  | 0.008 | 0.055 | 0.022 | 0.004  | 0.022  | 1.000 | 0.006    | 0.014       | 0.023      | 0.020  | 0.008        | 0.006   | 0.017    | 0.028 | 0.021 | 0.002 | 0.024  | 0.007     | 0.029 | 0.011  | 0.016 | 0.000 | 0.004 |
| *believe     | 0.016   | 0.000 | 0.016    | 0.005       | 0.012  | 0.009 | 0.004 | 0.000 | 0.010  | 0.032  | 0.006 | 1.000    | 0.015       | 0.006      | 0.008  | 0.000        | 0.014   | 0.009    | 0.009 | 0.004 | 0.011 | 0.014  | 0.009     | 0.014 | 0.005  | 0.010 | 0.007 | 0.002 |
| *government  | 0.023   | 0.001 | 0.005    | 0.002       | 0.008  | 0.009 | 0.013 | 0.006 | 0.009  | 0.002  | 0.014 | 0.015    | 1.000       | 0.013      | 0.010  | 0.006        | 0.002   | 0.000    | 0.004 | 0.004 | 0.009 | 0.007  | 0.004     | 0.014 | 0.019  | 0.009 | 0.000 | 0.012 |
| *political   | 0.016   | 0.000 | 0.017    | 0.002       | 0.000  | 0.004 | 0.115 | 0.013 | 0.018  | 0.004  | 0.023 | 0.006    | 0.013       | 1.000      | 0.002  | 0.011        | 0.011   | 0.061    | 0.004 | 0.042 | 0.004 | 0.021  | 0.012     | 0.000 | 0.005  | 0.010 | 0.009 | 0.005 |
| *point       | 0.015   | 0.001 | 0.008    | 0.005       | 0.005  | 0.003 | 0.026 | 0.017 | 0.011  | 0.010  | 0.020 | 0.008    | 0.010       | 0.002      | 1.000  | 0.019        | 0.013   | 0.006    | 0.013 | 0.004 | 0.002 | 0.009  | 0.002     | 0.007 | 0.005  | 0.007 | 0.004 | 0.009 |
| *interesting | 0.009   | 0.001 | 0.006    | 0.013       | 0.002  | 0.004 | 0.013 | 0.004 | 0.012  | 0.008  | 0.008 | 0.000    | 0.006       | 0.011      | 0.019  | 1.000        | 0.009   | 0.007    | 0.005 | 0.002 | 0.000 | 0.002  | 0.012     | 0.007 | 0.002  | 0.017 | 0.009 | 0.022 |
| *system      | 0.015   | 0.000 | 0.005    | 0.002       | 0.009  | 0.007 | 0.002 | 0.004 | 0.002  | 0.027  | 0.006 | 0.014    | 0.002       | 0.011      | 0.013  | 0.009        | 1.000   | 0.002    | 0.005 | 0.005 | 0.014 | 0.010  | 0.002     | 0.002 | 0.015  | 0.018 | 0.000 | 0.026 |
| *channel     | 0.017   | 0.001 | 0.024    | 0.006       | 0.000  | 0.007 | 0.035 | 0.004 | 0.004  | 0.000  | 0.017 | 0.009    | 0.000       | 0.061      | 0.006  | 0.007        | 0.002   | 1.000    | 0.000 | 0.026 | 0.000 | 0.002  | 0.037     | 0.002 | 0.005  | 0.005 | 0.014 | 0.000 |
| *life        | 0.029   | 0.000 | 0.003    | 0.002       | 0.009  | 0.005 | 0.010 | 0.006 | 0.004  | 0.022  | 0.028 | 0.009    | 0.004       | 0.004      | 0.013  | 0.005        | 0.005   | 0.000    | 1.000 | 0.009 | 0.040 | 0.010  | 0.005     | 0.005 | 0.008  | 0.015 | 0.002 | 0.010 |
| *tell        | 0.010   | 0.001 | 0.016    | 0.003       | 0.003  | 0.002 | 0.045 | 0.018 | 0.002  | 0.004  | 0.021 | 0.004    | 0.004       | 0.042      | 0.004  | 0.002        | 0.005   | 0.026    | 0.009 | 1.000 | 0.009 | 0.007  | 0.000     | 0.000 | 0.003  | 0.007 | 0.010 | 0.002 |
| *live        | 0.015   | 0.000 | 0.002    | 0.003       | 0.017  | 0.004 | 0.006 | 0.006 | 0.000  | 0.013  | 0.002 | 0.011    | 0.009       | 0.004      | 0.002  | 0.000        | 0.014   | 0.000    | 0.040 | 0.009 | 1.000 | 0.005  | 0.002     | 0.007 | 0.007  | 0.007 | 0.005 | 0.005 |
| *issue       | 0.012   | 0.001 | 0.005    | 0.005       | 0.009  | 0.007 | 0.030 | 0.002 | 0.012  | 0.020  | 0.024 | 0.014    | 0.007       | 0.021      | 0.009  | 0.002        | 0.010   | 0.002    | 0.010 | 0.007 | 0.005 | 1.000  | 0.012     | 0.002 | 0.003  | 0.013 | 0.005 | 0.005 |
| *politics    | 0.011   | 0.005 | 0.010    | 0.005       | 0.005  | 0.000 | 0.018 | 0.006 | 0.000  | 0.000  | 0.007 | 0.009    | 0.004       | 0.012      | 0.002  | 0.012        | 0.002   | 0.037    | 0.005 | 0.000 | 0.002 | 0.012  | 1.000     | 0.000 | 0.000  | 0.003 | 0.002 | 0.005 |
| *war         | 0.015   | 0.001 | 0.002    | 0.008       | 0.009  | 0.007 | 0.004 | 0.000 | 0.006  | 0.002  | 0.029 | 0.014    | 0.014       | 0.000      | 0.007  | 0.007        | 0.002   | 0.002    | 0.005 | 0.000 | 0.007 | 0.002  | 0.000     | 1.000 | 0.021  | 0.008 | 0.000 | 0.005 |
| *money       | 0.020   | 0.001 | 0.008    | 0.002       | 0.024  | 0.006 | 0.002 | 0.002 | 0.004  | 0.019  | 0.011 | 0.005    | 0.019       | 0.005      | 0.005  | 0.002        | 0.015   | 0.005    | 0.008 | 0.003 | 0.007 | 0.003  | 0.000     | 0.021 | 1.000  | 0.008 | 0.000 | 0.005 |
| *idea        | 0.015   | 0.004 | 0.008    | 0.007       | 0.007  | 0.004 | 0.012 | 0.008 | 0.011  | 0.012  | 0.016 | 0.010    | 0.009       | 0.010      | 0.007  | 0.017        | 0.018   | 0.005    | 0.015 | 0.007 | 0.007 | 0.013  | 0.003     | 0.008 | 0.008  | 1.000 | 0.000 | 0.003 |
| *lpf         | 0.002   | 0.006 | 0.008    | 0.003       | 0.004  | 0.006 | 0.006 | 0.008 | 0.004  | 0.009  | 0.000 | 0.007    | 0.000       | 0.009      | 0.004  | 0.009        | 0.000   | 0.014    | 0.002 | 0.010 | 0.005 | 0.005  | 0.002     | 0.000 | 0.000  | 0.000 | 1.000 | 0.003 |
| *part        | 0.013   | 0.000 | 0.005    | 0.003       | 0.007  | 0.006 | 0.006 | 0.004 | 0.013  | 0.007  | 0.004 | 0.002    | 0.012       | 0.005      | 0.009  | 0.022        | 0.026   | 0.000    | 0.010 | 0.002 | 0.005 | 0.005  | 0.005     | 0.005 | 0.005  | 0.003 | 0.003 | 1.000 |

## 8.14 – Appendix 14: IRC Interface Example (Quassel Client)



[illegible]

[illegible]

<dragos>

[illegible]

<dragos>

[illegible]

```
<dragos> sssssssssssssssssssssssssssssssssssssssssssss
```

<dragos>

[illegible]

## 8.16 – Appendix 16: Creative Use of Emoticons or “Smileys”

From various IRC channels, and *Wikipedia: List of emoticons*, which the reader would do well to visit if he does not understand the meaning of one or several of them. This is only a small peek and does not reflect the richness achieved with full use of *Unicode* characters. (See also [http://en.wikipedia.org/wiki/ASCII\\_art](http://en.wikipedia.org/wiki/ASCII_art).)

“Western-style” emoticons (font: Ubuntu Mono):

```
:) :-) :( :-(: :P :-P :D :-D D: D-: xP XP :o) :] :3 :c) :>
=] 8) =) :} :^) :つ) :-D :D 8-D 8D x-D xD X-D XD =-D =D =-3
=3 B^D :-)) >:[ :-(: :(-c :c :-< :つ C :< :-[ :[ :{ ;( :-||
:@ >:( :'( :'( :'-) :') D:< D: D8 D; D= DX v.v D-' :>:0 :-0
:0 :-o :o 8-0 0_0 o-o 0_o o_0 o_o 0-0 :* :^* ( '){' ) ;-) ;)
*-) *) ;-[ ;] ;D ;^) :-, >:P :-P :P X-P x-p xp XP :-p :p =p
:-p :p :p :-p :-b :b d: >:\ >:/ :-/ :- . :/ :\ =/ =\ :L =L
:S >.< :| :-| :$ :-X :X :-# :# 0:-) 0:-3 0:3 0:-) 0:) 0:^)
>:) >:) >:-) }:-) }:) 3:-) 3:) o/\o ^5 >_> ^ ^ <_< |;-) |-0 :-&
:& #-) %-) %) :-###.. :###.. <:-| 0_0 <*)-)-{ ><(((*) ><> \o/
*\0/* @}-;-'--- @>-->-- ~(_8^(I 5:-) ~:-\ //0-0\\ *<|:-) =:o] ,:-)
7:^] <3 </3
```

“Eastern-style” emoticons (font: Ume P Mincho):

```
^_ ^_ (>_<) (>_<)> (';) (^> (^_;) (-_;) (~_~) (°.°) (°_°) (°;) ^; ^_~;
(#.^#) (^_;) .. oO Oo. <コ: 彡 (°~)y-.. oO (-.~)y-°° (-_)zzzz (^_)(^_)-☆
((+_+)) (+o+)(°°)(°-°)(°°)(°_°)(°_>)(°_°) (o|o) <(^~)> (°o°) (^_)/ (^O^)/ (^o)
/ (^)/ (≥▽≤)/ (/●▽●)/ (^o)J ∩(・ω・)∩(・ω・) ^ω^ □ _(.)_ _(^)_ <(_)>
<m(_m)> m(_m) m(_)_m (°°)~ (^^)_U~~ (^^)_且~~ ☆彡 ☆彡 \(^o\)\ココハドコ?
(/口°)/アタシハダアレ? >)))彡 (Q)) ><ヨヨ(°)<<>)))彡 <°)))彡 >°))彡 <+))><<(*)
>=< ('_) (/;) (T_T) (;;) (;; (;:) (;O;) (:;) (ToT) (T▽T) ;;; ;;; ;n; ;; Q.Q T.T
QQ Q_Q (-_-)!! (-.-) (-_-)(- -) (;- -) C:。 三 (= =) ~>°) ~ ~ ~ ~° . _ ° ~
(=^·^=) (=^··^=) =^_ = (..) (..) ^m^ (·? (?_?) >^< <^!> ^/ (^*_*) §^.^§ (^<^)
(^.) (^ム) (^.) (^.) (^_) (^) (^) (^J) (*。*) (#.^#) (^-^) (^)/~~~ (^)/~
(;;)/~~~ (^)/~~~ ($··)/~~~ (@^)/~~~ (T_T)/~~~ (ToT)/~~~ ●~* (V)o ¥ o(V) \(^o~)
/ \(^o)/ \(-o-)/ \(^.) / \(^o)J (*0*) (*_*) (*_*; (+_+) (@_@) (@_@. (@_@;)

```

\(\odot o \odot) / ! (-\_-) / ~~~ピシー!ピシー! !(^)! (\*~)v (^~)v (^~)v (^▽^)(・▽・) ( '▽` )  
 ( ^▽^ ) ( ^v^ ) ( ' - ' \* ) ( ~o~ ) ( ~\_~ ) ( ~> ( p\_- ) ((d[-\_-]b)) (-"-) (一一`) ( ^\_メ ) ( \_-メ )  
 ( ^` ) ( ~\_メ ) ( -- ㄥ ) ( ・^・ ) < ~ ^ > < ^ ^ ^ > (一一;) (^0\_0^)(..)φ メモメモ  
 φ(..)メモメモ (●^o^●) (^v^) (^u^) (^◇^)(^o^)(^O^)(^o^)(^O^)^o^(\*^▽^\*)  
 (♣~^~) (□□; ) ° o° ° O° :O o\_O o\_0 o.O (o.o) (\*'▽`\*) (\*° ∇° )=3  
 ( ° 㐓° ) ( ° ◇° ) (\*~m~) \ (^ー^ ) 㐁 㐂 㐃 㐄 㐅 㐆 㐇 㐈 㐉 㐊 㐋 㐌 㐍 㐎 㐏 㐐 㐑 㐒 㐓 㐔 㐕 㐖 㐗 㐘 㐙 㐚 㐛 㐜 㐝 㐞 㐟 㐠 㐡 㐢 㐣 㐤 㐥 㐦 㐧 㐨 㐩 㐪 㐫 㐬 㐭 㐮 㐯 㐰 㐱 㐲 㐳 㐴 㐵 㐶 㐷 㐸 㐹 㐺 㐻 㐼 㐽 㐾 㐿  
 ..... φ(・▽・\*) キタ——(° ∇°)——!!!! \_|\_|O STO OTZ OTL orz (° 㐓° )  
 (㐓° □° )㐓 㐔 㐕 㐖 㐗 㐘 㐙 㐚 㐛 㐜 㐝 㐞 㐟 㐠 㐡 㐢 㐣 㐤 㐥 㐦 㐧 㐨 㐩 㐪 㐫 㐬 㐭 㐮 㐯 㐰 㐱 㐲 㐳 㐴 㐵 㐶 㐷 㐸 㐹 㐺 㐻 㐼 㐽 㐾 㐿  
 (ノ益 㐓)ノ彡 㐔 㐕 㐖 㐗 㐘 㐙 㐚 㐛 㐜 㐝 㐞 㐟 㐠 㐡 㐢 㐣 㐤 㐥 㐦 㐧 㐨 㐩 㐪 㐫 㐬 㐭 㐮 㐯 㐰 㐱 㐲 㐳 㐴 㐵 㐶 㐷 㐸 㐹 㐺 㐻 㐼 㐽 㐾 㐿

### Miscellaneous:

／人● 㐔 ●人\ (on #libertyinmind, font: Ume P Mincho)

㐔 㐕 㐖 㐗 㐘 㐙 㐚 㐛 㐜 㐝 㐞 㐟 㐠 㐡 㐢 㐣 㐤 㐥 㐦 㐧 㐨 㐩 㐪 㐫 㐬 㐭 㐮 㐯 㐰 㐱 㐲 㐳 㐴 㐵 㐶 㐷 㐸 㐹 㐺 㐻 㐼 㐽 㐾 㐿

(on #ubuntu-offtopic, font: ubuntu)