Computer–Assisted Conceptual Analysis of Textual Data as Applied to Philosophical Corpuses

Jean Guy Meunier
meunier.jg@gmail.com
Université du Québec à Montréal, Canada

Louis Chartrand
lochartrand@gmail.com
Université du Québec à Montréal, Canada

Mathieu Valette
mvalette@inalco.fr
National Institute for Oriental Languages & Civilizations
France

Jackie Chi Kit Cheung
jcheung@cs.mcgill.ca
McGill University, Canada

Marie-Noëlle Bayle
mrcal.bayle@gmail.com
Université du Québec à Montréal, Canada

Overview

Jean Guy Meunier and Louis Chartrand

Many DH projects call upon computer tools for descriptive and analytic purposes: lexicon statistics, concordances, parsers, descriptive statistics, classifications, topic modeling, annotations, automatic summarization, visualization tools, etc. They have been mainly applied to literary, political, journalistic, or otherwise mediatic corpuses. However, less work has been done on philosophical corpuses, or corpuses that have been tailored for the ends of philosophical investigation.

Computational approaches specialized for highly theoretical and abstract texts have proved their efficiency in various domains, particularly in those pertaining to the encoding, curation and presentation of textual data (e.g. digitization, web publishing, authorship analyses, etc.). These successes open avenues for more complex tools and methods to study linguistic features which are not directly observable, such as narratives, themes and concepts.

Concepts, in particular, constitute a key issue, given, on one hand, the polysemy of the concept of concept, and, on the other hand, its widespread use in philosophy and other disciplines. A conceptual analysis is a process through which we decompose and thus elucidate the meaning of a concept. While it is traditionally practiced from the armchair, concepts' meanings are reflected in the texts where they are expressed. As such, the development of methods and approaches to computer-assisted conceptual analysis of texts (CACAT) has the potential to make conceptual analysis more precise, more reliable, more exhaustive and more inclusive.

These new challenges call for an appropriation of modern computational tools which have demonstrated their potential at discovering such entities for natural language processing. The last two decades have seen important developments in computational linguistics and in machine learning which have enabled researchers to detect and manipulate various complex features of textual data. Complex objects such as entities, events, topics, arguments, or syntactic and discourse relations can now be detected and studied. Progress in modelling and learning approaches from fields like probabilistic modeling and neural networks have made it possible to represent complex representations between various latent and explicit textual features, and to learn them in efficient ways. Because they capture different aspects of concepts, including those which appear to be latent or implicit, the innovations could open new and exciting horizons for conceptual analysis.

In order to exploit this potential, digital humanists must participate in the conception of the aforementioned innovations. The humanities have problems and conceptual tools of their own, which differ from those of computer scientists, and, as such, ought to be enunciated and translated into tasks for algorithms to fulfill. On the other hand, computational approaches to conceptual analysis pose specific problems, which must be addressed as new methods are developed: indeterminacy of the nature of a concept and of the method of analysis, complexity in the relation between and among concepts, diversity of interpretations, incertitude in the evaluation schemes, limited set of computation tools to explore conceptual structures, shallowness of visualization tools, etc. These difficulties call for work from social scientists and humanists.

Our aim in this session is to give an overview of the challenges, avenues and opportunities that are shaping CACAT's development. Each paper plays a specific role within this session, so that information in one paper may serve as context for another. Jean-Guy Meunier's paper reviews the challenges facing CACAT, serving as context for the papers that follow. Mathieu Valette's paper (which will be moved to second place) expands on this topic. Drawing on the study of concept formation in philosophy of science, he offers a criticism on the concept of concept, proposes an alternative and draws the implications for the relationship between concepts and their concrete expressions in texts. In working to bridge the gap between the philosophical conceptual analysis in texts and modern techniques of computational linguistics, these two papers work from the conceptual analysis pole. The following two papers, by Louis Chartrand and Jackie Cheung, work from the computational linguistic pole, presenting models and techniques which have the potential of addressing the challenges of CACAT. The final paper, by Marie-Noëlle Bayle, is a recent application of CACAT that exemplifies its potential in discovering implicit dimensions to a concept.

Modelling Computer assisted conceptual analysis in text (CACAT)

Jean–Guy Meunier

Conceptual analysis paradigms

In many fields of scientific research, be they social sciences, natural sciences or even professional practices, abstract or highly theoretical concepts are explored to discover their content and deepen the knowledge they embed. However, there is no consensus on the nature of a concept or on the methodology to analyze them. For example, how would one proceed in analyzing the concept of Evolution in Darwin's writings? Three radically different paradigms parameterize the methodology: philosophical, linguistic and cognitive.

In the philosophical paradigm, concepts are identified to the meaning of predicative words. For some, their analysis aims at finding the conditions (necessary, sufficient, fuzzy, etc.) under which these words refer to objects, events or actions in a possible or actual world. For others, analysis consists mainly of identifying the sense or intention of these words as related to the epistemic or metaphysical conditions for their understanding. Finally, for some others, an analysis should consider the use and context (linguistic, social or other) of these words. Hence, in this philosophical paradigm, conceptual analysis becomes a sort of logico-pragmatic analysis of the meaning of words. In our Darwin example, this paradigm would therefore ask what are the meaning conditions of the word evolution when Darwin uses it.

In the linguistic paradigm, concepts are also related to the meaning of words. For the Saussurian structuralists a concept is the core meaning embedded in the structure of the signified (le signifié) of words. For the neo-structuralists, the generativists and the cognitivists, a concept is also equated to the semantic content of predicative linguistic expression, and meaning is understood as a complex set of semantic properties (features, relations, frames, nets, etc.,) underlying isolated words or their position in sentences and discourse. Here, conceptual analysis becomes identified with classical semantic analysis of words. In Darwin's works, the analysis would explore the semantics properties of the English word evolution: for instance, it would study its lexical content, its synonyms, its topics, is semantic nets, etc.

In the cognitive paradigm, concepts are the results of cognitive or mental operations. For psychology, they are seen as a sort of cognitive categorization. For the analytical and hermeneutic traditions of philosophy of mind, they are mental states or world representations. Conceptual analysis consists then in exploring how semiotic or linguistic forms embed categories, intentions, conceptual spaces, beliefs, mental states, Weltanschauung, etc. Hence conceptual analysis bears resemblance to an exploration of cognitive operations or states: representing, categorizing, reasoning, argumenting, entailing, etc. In our analysis of Darwin, this cognitive paradigm would focus the analysis on the mental operations underlying the meaning of evolution. How is this category of mental representation acquired, built reasoned on, argued, etc.?

Choosing a paradigmatic methodology for analyzing concept is difficult, then, because not one of them is canonical. Conceptual analysis becomes an even more acute problem when computations are introduced in the methodology. The level of complexity of the task is so high that is not obvious how a computer assisted conceptual analysis of text (CACAT) project can be realized. Should it be computer-tool-driven or model-driven?

Tool driven approaches

The first type of approach is tool-driven. Once a methodology inspired by one of the paradigms presented above is chosen, its practitioners use some computer programs already built and inserts them in

appropriate moments of the analysis procedure. Many computer tools for this task actually exist.

A first set of tools focuses on the lexical expressions of a concept. The most classical ones are concordancers, collocation and lexical analysers, taggers, etc. These tools explore the lexical properties and contexts of one or a few canonical predicates, expressing the specific concept to be analyzed. The limits of these types of tools lie in their underlying design hypothesis: a conceptual content is to be explored through specific canonical expressions. Such a hypothesis restricts the exploration of the conceptual content to one or a specific number of predicates. This is problematic, for as we know, concepts can be expressed in language in a myriad of ways. For example, it would be very problematic to restrict Darwin's concept of evolution to the analysis of the word evolution alone. Secondly, they may produce results that are larger than the original text. This is the case of the concordance of the concept of Esse in the Thomas Indexicus. Finally, sometimes, the opposite happens. These tools may deliver only a fraction of the overall textual segments or word collocations whose content is pertinent. For instance, in Darwin only uses a few dozen times the lexical form evolution. Hence concordance, collocation, etc. on such a small sample are not very fruitful for a conceptual analysis.

A second set of tools highly influenced by classical AI approaches focuses on natural language processing (NLP). These tools are sensitive to various meaning aspects of words, such as their semantic definition, their encyclopedic, pragmatic discursive content, etc. They promise to deliver finer results for a conceptual analysis. But these tools also have limits. Their underlying hypothesis is that these semantic, pragmatic and encyclopedic information added in the grammar and the lexicon will enhance the exploration of conceptual content. Unfortunately, the added information has often been collected from common and ordinary semantic knowledge of shared language usages. Such tools will then often tend to identify already known properties belonging to this common information about the lexical conceptual word under inquiry. And most of the time, it will ignore the properties that precisely are the one that are specific to the concepts analyzed mainly when they are original, and belong to a reflexive, creative literary or reflexive discursive process, etc. These semantic properties would not be part of the common doxastic conceptual content. For instance, a philosophy scholar using such types of tools would not be very

satisfied in discovering that Darwin's concept of evolution is a name meaning an action of the type change and applied to the object: natural species.

Recently, a last set of tools that are more mathematically grounded, such as neural net and Bayesian classification, vector semantics, machine learning, deep learning, etc., have become appealing and are used in language processing, They can process large data and learn semantic information by themselves. But like the other set of tools they have their limits. First, they are nor readily usable. They are in fact very complex algorithms, and are not easily mastered by humanities scholars. Secondly, their lack of traceability becomes a major obstacle when applied to large and theoretical textual data where results become difficult to evaluate. Thirdly, they seem more successful for information retrieval applications than for digging into deep conceptual content. For the moment, we are not sure that how they can effectively assist conceptual analysis.

From these remarks, it does not seem to us that conceptual analysis can only be a serendipity tool-driven approach. The results produced by these tools have not yet convinced the scholarly community that practices expert conceptual analysis.

Model driven approaches

The second type of approach is model-driven. Recent philosophers of science such as Morgan and Morrison (1999), Giere, (1999), Leonelli (2007) for instance, see science as a building models process where models are heuristic means for describing, explaining and understanding reality. And Mc Carthy (1999), has seen this modelling approach as a means of better understanding digital humanities interpretative projects. For our part, we explore this hypothesis and see CACAT as type of scientific inquiry where various models are used as intermediaries for understanding the analysis of highly theoretical and abstract concepts. In this perspective, we distinguish four types of models: conceptual, formal, computational and experimental.

A conceptual model defines parameters for identifying, explaining and understanding the properties and structure of linguistic items expressing conceptual content. A formal model translates certain aspects of a conceptual model in some controlled formal language that describes or identifies properties and relations of these conceptual expressions. A computational model translates some formal expressions of the formal model into algorithms and programs. Finally, an experimental model designs implementation of these formal models in a concrete computer where the analysis can

be simulated and ultimately evaluated in correspondence to the other models.

In a concrete procedure, all these models interact and can be modified and adjusted. This allows the inquiry to be controllable and repeatable. It has been our own experience that, if a computer assisted conceptual analysis project is to be successful it must construct at least these four models. A CACAT project cannot bypass these models and their interactions.

Designing these models, their interactions and their experimentation to see CACAT as a scientific endeavour and not just computer gadget exploration. But each model is not built easily. And nothing comes smoothly. They are part of the research process. And much work must be done to clarify the conceptual, formal, computational and experimental models pertinent for a successful and pertinent conceptual analysis.

Digital epistemology for concept analysis

Mathieu Valette

In the humanities, theory is most of the time outlined with texts: papers, books, conference presentations, lectures etc. we claim that the scientist is first a reader and a text producer. This textuality is so ordinary that it is almost invisible, and, as such, not considered as an object of science. Moreover, theories are read as synchronic systems, or even achronic systems, depending on their specific purposes (describing one fact, explaining one phenomenon...). Scientists appropriate models and concepts like tools; they have to know their function and how to manipulate them, but they do not care about knowing practical details of their enunciation. In fact, they ignore them, more or less. They find such details embarrassing, because they make concept borders fuzzy: lexicons, glossaries, and also handbooks, as they extract the concepts from their context, and standardise the definitions, creating an illusion of stability and tangibility. But concept textuality necessarily has an incidence, not only on interpretation, but also on theorisation. If the scientist is a text producer, then theorisation is the construction of meaning. Theorisation is forced by enunciation, and scientific works, beyond their materiality, can be considered as text.

The textual aspect of scientific works had been noticed by those in Europe looking at epistemological culture. In this respect, French philosopher Michel Foucault's works, in the 1960s, must be acknowledged (see e.g. Foucault 1969). Foucault put in place a philo-logical analysis of discourse centred on the combination and evolution on specific discursive structures. His purpose is, firstly, to recognize "discursive formations", i.e. stabilized relations, regularities between objects, types of speech act, concepts and topics; and, secondly, to recognise breakpoints in idea system history. Foucault followed the example of some of his famous predecessors, such as Gaston Bachelard, Georges Canguilhem and Martial Gueroult. Bachelard's notion of Epistemological break, or Canguilhem's notion of concept shifts shows, for instance, that the history of a concept is not that of its increasing rationality and refinement, but that of the different fields in which they have been designed and validated. What we will call digital epistemology is a linguistic approach to this style of French epistemology.

Our topic is the study of scientific texts using, on the one hand, corpus linguistics tools which have been developed over the 40 last years and, on the other hand, a linguistic methodology (see Rastier 2009, Valette 2003). Thus, our purpose is to develop tools and methodology Foucault did not have, among other reasons because some textual phenomena—as, for example, lexicon evolution, which depends on the reader's subjectivity—are invisible to a classical philological analysis. Concept emergence, concepts' individual and inter-related evolutions, the appropriation of a specific thematic, palinode, etc. constitute further examples. We do not adopt the logician's position, considering that conceptualization is a linguistic phenomenon with its own construction rules linked to a particular function of language. Neither do we ignore the psychological, social and interactional reasons of the development of concepts. Firstly, we consider that textuality—i.e. the constraints of the textual layout, formulations, be they constraints of syntactic, semantic, lexical or related discursive traditions (including genres and speech)—plays a major role in concept formation. Secondly, we do not consider texts only as resources to mine and extract terminological and conceptual material, but as archives, or, in other word, as the objective tracks of the process of creating concepts.

In essence, we focus here on concept emergence considered as the result of a slow and gradual stabilisation of contextual semantic feature. Drawing on recent critical readings of Saussure's semiology (see Rastier 2015), we propose to consider a concept as a stabilized semantic form; that is, as a combination of semantic features (or semes) mainly inherited from various contexts in which it has occurred. Eventually, we link concept design with text production rather

than identification of items in a general ontology (Valette 2010).

Topic models for conceptual analysis

Louis Chartrand

The last two decades have seen the rise of topic models in natural language processing (NLP). From the early successes of Latent Semantic Analysis, which decomposes datasets into "conceptual" dimensions, the introduction of probabilistic and generative models have enabled the discovery of underlying structures that condition the lexicon of a text. Those structures, in turn, are used to construct meaningful representations of corpuses and documents, and have proven fruitful in improving performances in many NLP tasks.

Those tools have interesting potential for the Digital Humanities, as they discover entities which are, on one hand, robust features of textual data, and, on the other hand, easily representable and interpretable by humans. For instance, topics may help in tasks such as categorizing documents or selecting a relevant sub-corpus for analysis. However, once topics are represented using the words to which they are likely to be associated, they can also be used to make sense of what a set of textual segments are about, or to visualize the evolution of discourse in a corpus through time. As such, topics have interesting potential when it comes to representing textual data and improving our analyses of it.

In this presentation, some prominent topics models—LSA, LDA and DTM—will be presented and contrasted, and their potential uses for Digital Humanities will be discussed.

Latent Semantic Analysis (LSA)

Introduced by Deerwester et al. (1990), LSA used tools from linear algebra, in particular singular vector decomposition, to transform the representation of text segments in the form of word counts to a representation in the form of participation to "concepts", or semantic dimensions.

As words give us a good idea of what a text is about, it is common practice in text mining to represent text segments by counting its words. A document containing "apple", "orange" and "pear" a high number of times each is likely to talk about fruits. And if multiple documents share the same words, they are likely to share common topics. However, this approach does not fare well with synonyms, which it does not recognize.

The LSA uses co-occurrence in word uses to synthesize the word-count representations into more compact semantic dimensions. As synonyms tend to have the same cooccurents, they also tend to participate to the same semantic dimensions. As a result, the new representation is closer to a semantic representation than was the word-count representation, hence the name "Latent Semantic Analysis".

Latent Dirichlet Allocation (LDA)

While LSA still is part of every NLP reputable toolset, it falls short in at least two key aspects. Firstly, its semantic dimensions are hard to read for a human: from a list of its most prominent words, it is usually hard to give a satisfying interpretation of what a semantic dimension is about (Chang et al. 2009). Secondly, it has no clear hypothesis as to how text is structured. On one hand, this makes it harder to explain semantic dimensions in terms of linguistics, psychology or discourse analysis. On the other, it means that LSA gets only part of the picture, and better algorithms with additional assumptions might produce better semantic dimensions.

Latent Dirichlet Allocation (Blei et al. 2003) is a probabilistic models which attempts to address this latter issue, and ends up addressing the former as well. It supposes that in a corpus, there is a certain number of topics, which, when activated, make it more or less likely for specific words to be present. Thus, when someone writes a document, LDA assumes that she selects a certain restricted number of topics, which in term condition which words will be found in the document. Using this assumption and an arbitrary number of topics, the algorithm infers the most likely list of topics, and their most likely assignments to documents.

As such, it produces once again a representation of documents or text segments from word counts, but in terms of topics rather than more abstract conceptual dimensions. The words most likely to be present when a given topic is activated are often visibly related, either semantically or because they participate in a transparent narrative. As such, they are easily read by a human interpreter, and can be used to give a sense of what documents, sub-corpuses or textual segments are about.

Dynamic Topic Models (DTM)

Another boon of LDA is that its probabilistic model can be modified account for particularities of the corpus, or to model features that we want to study in particular. For example, if we have a corpus that spans

across decades, we might expect topics to evolve with time, as society and culture change.

To model this, the algorithm devised by Blei & Lafferty (2006) uses a corpus split in time slices (say, per year) and topics are split accordingly, such that topic 1 at time 1 is different from topic 1 at time 2. Then, a Markov assumption is enacted on the time series: topic 1 at time 1 conditions topic 1 at time 2, which conditions topic 1 at time 3, etc. This gives topics the freedom to evolve, while enforcing a certain degree of conservatism.

Using this, one can not only track topics more efficiently, but also see the evolution of topics across time.

What is a topic?

What is it, however, that we are talking about when we speak of LSA's conceptual dimensions or LDA's topics? Can it be equated with the notion of TOPIC that we encounter in discourse analysis, for instance?

While there are a variety of definitions for words such as "topic" and "theme", most agree that a topic is what a text is about (Rimmon-Kenan, 1995). On this score, LDA's topic does seem to agree with the common notion of TOPIC: a word list representing a LDA topic is read as a representation of what the textual data is about (Blei, 2013). Furthermore, the information the probabilistic model captures is the one that is redundant across a number of text segments. As such, it highlights words and concepts which keep coming back as they put in relation with various entities in sentences. In other words, textual discourse and narratives are being sewn around them.

On the other hand, humans tend to make slightly different representations of topics compared to machines (Chang 2010), more readily constructing topics around concepts and thus providing sparser (more compact) representations. As Chang suggests, this might be because humans build these representations using general domain knowledge, whereas topic models try to infer this knowledge from word distributions. This seems to tell us that we should understand LDA topics as indicators, or reconstructed traces, of the topics that underlie a text, but not as true representation of topics themselves.

Using topic models in conceptual analysis

As Chang's 2010 experiment suggest, topics entertain special relation with concepts, as a topic tends to be associated with a restricted number of concepts which are expressed very often in the text.

As such, topic models' potential in representing textual data can be exploited to discover associations that are likely to be useful to conceptual analysis and other philosophical analyses. For instance, it can help the analyst identify the most important parts of a corpus, and those that can be discarded. They can also be leveraged to build representations of the contexts in which the concept of interest appears, thus giving a sense of the topics with which it is associated. Using DTM, one can also get a sense of the evolution of a concept within a diachronic corpus. Beyond discovery of new informations concerning a concept's expression in a corpus, topic models can be useful to test some association, as the structures they uncover are relatively robust.

That said, as the DTM model shows, topic models can be used in a large variety of use cases, as their model can be expanded to take into account a corpus' metadata and thus open new and innovative avenues for conceptual analysis and the Digital Humanities in general.

Unsupervised natural language processing for conceptual analysis of events
Jackie Chi Kit Cheung

In unsupervised machine learning, an algorithm is trained to discover regularities in data without access to human-provided labels. Such techniques can be useful in conceptual analysis of text, in cases where we do not have or want to impose a schema on the text corpus under analysis. The basic intuition behind unsupervised natural language processing techniques is that objects that appear in similar contexts in the data should be assigned similar representations, such that they can be grouped into clusters.

Unsupervised models differ according to several characteristics, from the type of information that is made available to the learner, to how similarity is defined between the different objects that are modeled, to the expected form of the output cluster that is learned. For example, the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) is a probabilistic model which is given access to multiple documents for training. The crucial assumptions behind the LDA model are that each document can be described as a mixture of multiple topics, and each word in a document is generated by one of the topics in that mixture. As a result of training an LDA model, multiple topics are learned, which correspond to clusters of words that tend to co-occur in the same documents.

More recently, there have been a number of unsupervised models that have been used to discover the structure of a sequence of entities and events that appear according to some narrative in the natural language processing literature (Chambers and Jurafsky,

2008; Cheung and Penn, 2013). This is accomplished by explicitly modelling the sequential dependencies of events as they appear in a document. I will provide an overview of the assumptions of the event structure being learned by such models. For example, some methods produce discrete sequences of prototypical event and participant roles. In the work of Chambers and Jurafsky, (2008), narrative chains are learned corresponding to prototypical roles in a narrative. A chain such as _ accused X; X claimed _; X argued _; _ dismissed X might correspond to a defendant in a trial. Other work frame the problem as a task for probabilistic learning. Cheung and Penn (2013) define a probabilistic sequence model, in which the structure of an event and its participants are explicitly represented in the model as latent random variables. The nature of a learned cluster, then, would be how it influences the conditional probabilities of generating other cluster labels, as well as the word emission distributions from that latent topic (as in an LDA model).

I will discuss how such models can be used to discover templates of prototypical events, including how events and event participants are typically expressed in language. Such approaches can easily be applied to multiple domains, including texts in the legal or medical genres, because they make minimal assumptions about the structure of events, and do not require training data. I also discuss other applications of these models to information ordering, and automatic summarization, which may be of interest to researchers in conceptual analysis for the digital humanities.

A computer–assisted analysis of SYMPTOM in psychiatry

Marie–Noëlle Bayle

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) is a general classification and diagnostic tool used in the rich and diverse universe of mental health. Being widely distributed and available online, it allows everyone to have a direct access on how to make a psychiatric diagnosis. To facilitate its reading, laypeople and professionals alike may consult definitions for important notions in the glossary section. However, these few lines will often fail to capture the complexity of a term. For instance, at the core of the clinical assessment of a disorder lie its signs and symptoms. Therefore, a proper understanding of what a symptom means, and how this concept relates to the disorder, is essential to the diagnostic approach.

In the DSM-5 glossary, symptom is defined as "A subjective manifestation of a pathological condition."

In the practice, it is often treated as a necessary and/or sufficient condition for to seek diagnosis, or as a constraint on possible diagnoses. As such, non-doctors and patients often think of a symptom as a singular event, and conceive of it purely in terms of content and a means to a diagnosis.

However, as experience of mental disorders is often messier, we may wonder if such a notion of SYMPTOM overlooks important features of the way this notion is actually reflected in the DSM-5. Our hypothesis is that both the glossary definition for this concept and the common understanding of this notion fail to account for such essential aspects. To provide evidence for this claim, we performed a computer-assisted conceptual analysis of text (CACAT) for the concept SYMPTOM in the DSM-5. We find evidence that SYMPTOM is strongly associated to a dimension of temporality, and that it is expressed in relation not only with the disorder, but also remission. As such, it is proposed an improved definition would not only better reflect the content of the DSM-5, but could also contribute in a better understanding of assessment and practice of diagnostic.

Method

The dataset consists in a small corpus composed with the most relevant chapters of DSM-5. Computational Text mining method and manual qualitative approach were used in a processing chain for the conceptual analysis of SYMPTOM. Firstly, extraction of the textual data from noisy sources and cleaning of the text were performed. Secondly, all sentences with the term "symptom" in it, plus one sentence before and after, were extracted, yielding a set of textual segments, on which stemming and lemmatisation were performed. Thirdly, the pretreated data was used to create a document-term matrix with the TF-IDF weighting scheme. Fourthly, from the matrix, textual segment clusters were produced with the k-means algorithm. Fifthly, the most salient words in each cluster and in the whole subcorpus were first represented using word clouds. Finally, relations of similarity between the most relevant words in each cluster and in the subcorpus were represented in a 3d space. All steps were performed using common R modules (tm, RWeka, qdap, cluster, knn, ggplot2, rgl). Each cluster was interpreted as a specific field in which a hypothetical conceptual property of SYMPTOM is expressed. Categorization was done by annotating manually the most typical textual segments in every cluster according to cosine distance to the centroid. Annotations consisted in the main conceptual property of SYMPTOM expressed in a segment.

Syntheses of these annotations were done for each cluster.

Experimentation

From the subcorpus, 2036 sentences containing the term "symptom" were extracted which contained 5761 different word types. The words most associated with symptom in this subset of the corpus are disord (disorder), criteria, presen (presence), sever (severity), medic (medical, medication). Using k-means, 30 clusters were extracted but 17 are deemed noisy, as they contain 10 textual segments or less. Most of the remaining clusters are located in the section II of the DSM (diagnostic criteria and codes), several having most of their segments in a specific disorder chapter. The converse, however, does not hold.

Discussion

Analysing those clusters reveals that SYMPTOM has a transdiagnostic property, and is not only defined by its specific content. For example, let us examine cluster #8, which contains 98 segments, 90% of which fall in the three chapters about psychotic and mood disorders. Symptom in this cluster is linked with the words depress, hypomania, mania, but also with episode, period, full, meet. Furthermore, annotation of joined documents shows that temporality is a conceptual property of SYMPTOM. It modifies the pathological dimension of its content. SYMPTOM is not in direct causal relation with DISORDER; it is a dynamic sign whose presence needs to be situated in an episode, regardless of whether its content is depressive or manic. Therefore, the mere presence of a symptom is not sufficient for a diagnosis. Conversely, SYMPTOM is also linked with the concept of (partial) remission. SYMPTOM and DISORDER have a complex relationship on a continuum between negative (remission) and positive (disease) poles.

In conclusion, a mixed method, combining computational and manual processing and using quantitative and qualitative approaches, was applied in our conceptual analysis of the concept SYMPTOM. SYMPTOM appears to be a more complex and dynamic concept than patients and other non-doctors usually understand it to be. As a result, a better understanding of this complexity would likely profit assessment, diagnosis and treatment of mental disorders.

Bibliography

American Psychiatric Association, (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association.

Bachelard, G. (1938) La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective. Vrin, Paris; The Formation of the Scientific Mind. Clinamen, Bolton, 2002.

Blei, D. M. (2012) Probabilistic Topic Models. Communications of the ACM 55, 4:77.

Blei, D. M. (2013) "Topic Modeling and Digital Humanities." Journal of Digital Humanities, April 8, 2013.

Blei, D. M., and Lafferty, J. D. (2006) Dynamic Topic Models. In Proceedings of the 23rd International Conference on Machine Learning, 113–120. ACM.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research 3: 993–1022.

Chambers, N., and Jurafsky, D. (2008) Unsupervised Learning of Narrative Event Chains. ACL, pp. 789-797.

Cheung, J. C. K., and Penn, G. (2013) Probabilistic Domain Modelling With Contextualized Distributional Semantic Vectors. ACL, pp. 392-491.

Chang, J. (2010) Not-so-Latent Dirichlet Allocation: Collapsed Gibbs Sampling Using Human Judgments." In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 131–138. CSLDAMT '10. Association for Computational Linguistics, Stroudsburg, PA, USA.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L. and Blei, D.M. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models." In Advances in Neural Information Processing Systems, pp. 288–296.

Chartier, J. F., and Meunier, J. G. (2011). Text Mining Methods for Social Representation Analysis in Large Corpora. Papers on Social Representations, 20: 37.1-37.46.

Danis J. and Meunier, J. G. (2012). CARCAT : Computer-Assisted Reading and Conceptual Analysis of Texts : An experiment applied to the concept of evolution in the work of Henri Bergson. Digital Studies/Le champ numérique, 3(1).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R.A. (1990) Indexing by Latent Semantic Analysis. JASIS 41, 6: 391–407.

Foucault, M. (1969) L'archéologie du savoir. Gallimard, Paris; The Archaeology of Knowledge (1969), Routledge, 1972.

Giere, R., (1999), Using Models to Represent Reality, in: L. Magnani, N. Nersessian and P. Thagard, (eds.), Model-Based Reasoning in Scientific Discovery. Plenum Publishers, New York, pp. 41–57.

Hofmann, T. (1999) Probabilistic Latent Semantic Analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 289–296. Morgan Kaufmann Publishers Inc.

Leonelli, S. (2007) What is in a Model? Combining Theoretical and Material Models to Develop Intelligible, Modeling Biology: Structures, Behaviors, Evolution. In Manfred Dietrich Laubichler, Gerd B. Mülle (eds.) Modeling Biology. MIT Press.

McCarthy, W. (2004). Humanities Computing. Palgrave MacMillan.

Morgan, M.S., and Morrison, M. (eds.), (1999), Models as mediators. Perspectives on natural and social science. Cambridge University Press.

Rastier, F. (2001) Arts et sciences du texte. PUF, Paris.

Rastier, F. (2015) Saussure au futur. Les belles lettres, Paris.

Rockwell, G. (2003). What is Text Analysis, Really? Literary and Linguistic Computing, 18(2).

Salton, G., and McGill, M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Turenne, N. (2016). Analyse de données textuelles sous R. ISTE Éditions, coll. Sciences Cognitives. Londres.

Valette, M. (2003) Conceptualisation and Evolution of Concepts. The example of French Linguist Gustave Guillaume, Academic discourse – multidisciplinary approaches, Kj. Fløttum & F. Rastier, eds., Novus Press, Oslo, pp. 55-74.

Valette, M. (2010) Des textes au concept. Propositions pour une approche textuelle de la conceptualisation », Actes des 21es Journées francophones d'Ingénierie des Connaissances (IC'2010) (8-11 juin 2010), Nîmes Sylvie Despres, éd., Publication de l'Ecole des Mines d'Alès, pp. 5-16.

Widdows, D. (2004). Geometry and Meaning. CSLI Publications, Center for the Study of Language and Information, Leland Stanford Junior University