Reading Darwin between the lines: a computer-assisted analysis of the concept of evolution in the *Origin of species*

Maxime Sainte-Marie, Jean-Guy Meunier, Nicolas Payette,
Jean-François Chartier

[1] Laboratoire d'Analyse Cognitive de l'Information (LANCI) – Université du Québec à Montréal (UQAM)

Abstract

While Darwin's first major work, *On the Origin of Species by means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*, is generally considered as the birth document of the theory of evolution, studies on and around this book often overlook the fact that the word evolution itself is rarely used by the author and thus complicates any attempt to properly analyze the concept. This specific issue will be addressed via a computer-assisted analysis of the concept of evolution in the *Origin of Species*: by means of automatic classification, categorization, and annotation strategies used in text mining, a new and automated approach for "reading Darwin between the lines" is here proposed, aiming to give a full account of the said concept regardless of any proper designation.

Keywords: Text mining, Conceptual Analysis, Darwin, Evolution, Philosophy of Biology

1. Introduction: the two meanings of "evolution"

Simultaneously praised and condemned by both clergymen and scientists, the Darwinian theory of modified descent by means of natural selection marks the birth of a radically new and modern conception of nature, life, science and man, built on the revitalization and reworking of an old biological concept: evolution. Whereas Darwin is nowadays considered the founder of the modern theory of evolution, he wasn't however the first to use the word in a biological context: as noted a while ago by Thomas Henry Huxley in his 1878 article on evolution published in the *Encyclopaedia Britannica*, the word "evolution" has been historically used by naturalists and biologists for two different purposes: "initially, to refer to the particular embryological theory of preformationism; and later, to characterize the general belief that species have descended from one another over time" (Richards, 1992: 4). At the time Darwin published the first edition of *On the Origin of Species by means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* (1859), both embryological and specific uses of the term "evolution" were still very much in use by supporters and opponents of transmutation theory alike.

1.1. Embryological evolution

Deriving from the Latin *evolutio*, which refers to the scroll-like act of unfolding or unrolling, the word «evolution» first occurs in a biological context in 1670, in an anonymous paper published in the *Philosophical Transactions of the Royal Society*. Reviewing Jan Swammerdam's (1637-1680) book *Historia Insectorum Generalis*, this article explained that by applying the word

change to insects, the Dutch entomologist meant nothing else "but a gradual and natural evolution and growth of the parts" (cited in Bowler, 1975: 97). In the following century, the word is used by the Swiss anatomist Albrecht von Haller (1708-1777) to refer to the preformationist theory, according to which embryos are miniature adults pre-existing conception and whose development or "evolution" during gestation implies some form of expansion or swelling caused by the absorption of nutrients. Haller also observed that by 1750, this theory of evolutions split into two rival interpretations, an ovist one, according to which these miniature adults are located and develop in the female's egg, and a spermist one, originating from the works of Hermann Boerhaave (1668-1738) and stating that the embryos reside and "unfold" in male semen (Richards, 1992: 5). Charles Bonnet, the main proponent of the ovist theory, initially used the word "*développement*" to refer to this process, but later contributed to the use of the word "*évolution*", notably through his work *Considérations sur les Corps Organisés*.

Experimental and theoretical objections to these theories by John Needham (1713-1781), Georges Leclerc, comte de Buffon (1707-1788) and Caspar Friedrich Wolff (1734-1794) rapidly brought epigenetic theories to prominence, according to which embryos and organs are formed out of undifferentiated matter sequentially, through the agency of mechanistic or vital principles or forces. Increasingly popular among the leading biologists of the early nineteenth century, this idea of epigenesis would receive its fullest and most modern experimental and theoretical account in the works of Karl Ernst von Baer's. In his 1828 book *Über Enwickelungsgeschichte der Thiere*, von Baer universally characterized embryological development as a gradual differentiation process leading from homogeneous matter to the production of heterogeneity and complexity of structure, process he usually referred to using the word *Entwickelung*, followed by the Latin *evolutio* in parentheses. However, mild these parenthetical additions may seem and despite radical shifts in theory and meaning, the ground-breaking importance of von Baer's work and its diffusion in the scientific community through numerous translations, commentaries and appropriations significantly contributed to consecrate the use of the word "evolution" for embryological purposes.

1.2. Specific evolution

Even though the concept of evolution as species progression emerged in the beginning of the 19th century, allusions to species change are rather old. Anaximander, Anaxagoras and Aristotle all wrote about the emergence of new kinds of animals, and Empedocles (fl. 444 B.C.) even proposed a simple model of natural selection, according to which accidental mutations in creatures might improve their chance of survival. Similar conceptions can also be found in the writings of later philosophers like Francis Bacon, de La Mettrie, Diderot and Kant. In early biology, Buffon and Linné (1707-78) respectively invoked environmental changes and hybridization as sources of species creation, while Erasmus Darwin (1731-1802) presented in his *Zoonomia* a rather original and dreamlike theoretical account of species alteration.

The first thorough model of specific metamorphosis however came from Jean-Baptiste de Lamarck (1744-1829). In his theory, first sketched in 1800 and fully developed in his *Philosophie Zoologique* (1807), environmental changes, on the one hand, and adaptive and heritable behaviour on the other, both contributes to modify animal parts and thus leads to species transformation over time and generations. Lamarck himself never used the word 'evolution' to refer to this process, but his commentators, detractors, readers and followers often did: Étienne Renaud Serres used the expression *théorie des evolutions* in his 1827 article *Théories des formations organiques* to refer both «to the recapitulational *métamorphoses* of organic parts in

the individual and the parallel changes one sees in moving (intellectually) from one family of animals to another and from one class to another» (p. 83); in his *Principles of Geology* (1830-1833), Charles Lyell used the word "evolution" to refer to and reject Lamarck's transmutation theory; von Baer (1792-1876), in his rejection of transmutationism and recapitulationism (the idea that the embryo goes through the development stages of the species), used the word "evolution" to refer to both processes. In light of these examples, as well as those found in the writings of, Joseph Henry Green (1791-1863), Robert Grant (1793-1874) and Richard Owen (1804-1892), it is obvious that «by the 1830s, the word "evolution" had shifted 180 degrees from its original employment and was used to refer indifferently to both embryological and species progression» (Richards, 1992: 15).

Further theoretical and experimental investigations contributed to further establish this dual use of "evolution": in the third edition of his *Principles of Physiology* (1851), William B. Carpenter used "evolution" not only to refer to von Baer's principle of embryological development, but also when noting that the same principle can be traced out in the fossil record. Even though Carpenter was opposed to the idea of transmutation, his dual use of "evolution" had an obvious influence on the theory of the younger Herbert Spencer (1820-1903): in his 1852 essay "The development hypothesis", Spencer openly supports transmutation and refers to it as the "Theory of Evolution", while pointing «to embryological "evolution" as an illustration of the ability of organic structures to modify themselves» (Bowler, 1975: 106).

In light of these historical considerations, the dual use of "evolution" was still in effect at the time Darwin first released the *Origin of Species*. However, whereas the most influential and renowned biologists of the time all seemed to endorse this "dual" reference to embryological and specific development, little is known about Darwin's own stance in this matter. Did he also use the word "evolution" to refer both to embryological and specific development?

2. "Evolution" in the *Origin of species*: lexical and conceptual issues

Finding out if Darwin's use of the word "evolution" encompasses both embryological and specific development is not an easy task: while the *Origin of species* is generally considered as the birth document of this theoretical revolution, studies on and around this book often overlook the fact that the word *evolution* itself is rarely used by Darwin. In the first (November 24th 1859), second (January 7th 1860), third (March 1861) and fourth (June 1866) editions, there is only one occurrence related to the term evolution: it is the last word of the conclusion of the work, *evolved*. In the fifth edition (1869), the same term *evolved* appears a second time, the first occurrence appearing in the fourteenth chapter and the second at the same last spot as in the earlier editions. Surprisingly, only in the sixth and last edition (1872) are the term *evolution* and its derivatives more extensively and systematically employed (for a more precise list of occurrences, see Tab. 1).

This lexical scarcity of the word "evolution" and its derivates doesn't necessarily mean however that Darwin didn't speak of evolution at all. In fact, in earlier studies of the concept of "evolution" as well as in the current debate on the matter, it is never easy to tell whether one refers to the concept of evolution or the term that refers to it. This might come as a surprise, considering the functional and cognitive differences between the operations of conceptualization and lexicalization. To better distinguish these two processes as well as to clarify the status and meaning of "evolution" in the *Origin of Species* and its many interpretations, a quick look at what distributional semantics tell us about words, concepts, and their relations might seem here necessary.

1st Edition (1)	*evolved*: XV (490)
2nd Edition (1)	*evolved*: XV (490)
3rd Edition (1)	*evolved*: XV (525)
4th Edition (1)	*evolved*: XV (577)
5th Edition (2)	*evolved*: XIV (573), XV (579)
6th Edition (14)	*evolution*: (VII: 201(2), 202), VIII (215), X (282), XV (424 (3))
	evolve: VII (191)
	evolved: VII (191, 202(2)), XV (425, 429)

Table 1: Occurrences of "evolution" and its derivatives in the Origin of Species

According to distributional semantics, «meaning is more easily stated as a property of word combinations (or of words in combination) than of words by themselves» (Harris, 1991: 325). This idea rests the correlation of redundancy with information, as pointed out by Shannon and Weaver's mathematical theory of communication: each word of a sentence brings its own constraints to the whole, reduces the sets of possible words that could fit in the sentence, therefore increasing the total information conveyed and structuring the semantic dimension of the sentence. In this sense, meaning doesn't reside in words *per se*, but "between" them, that is, in the underlying word combination networks they refer to and that constitute their conceptual dimension. Meaning of words is thus conceptually distributed in the sense that «similarities and patternings among the co-occurrence likelihoods of various words correlate with similarities and patternings in their types of meaning» (Harris, 1991: 341).

Such conceptual and lexical considerations, while emphasizing the distinction between the semantic associations of specific concepts and their embodiment in natural language, also seems to imply the possibility of "reading between the lines", that is, of identifying and analyzing concepts on the sole basis of their relations with other words and concepts and independently of any proper lexicalization. In view of this, the fact that the word "evolution" itself is rarely found in the *Origins of species* doesn't necessarily imply that the inferential structure it refers to and that constitutes its conceptual dimension isn't present elsewhere in the text and can't be studied in its stead. In this sense, taking into account word combinations similar to those where the word 'evolution' occurs instead on focusing solely on the latter might be the most reliable way to determine whether or not Darwin's concept of evolution in the *Origin of Species* refers to both embryological and specific development, like most biological theories and works of the same period. However, since dealing with word combination networks and statistics might be hard to deal with manually, the use of text mining strategies might seem here natural, even necessary. In view of this, a new computer-assisted conceptual analysis methodology is proposed here, one which aims to explore and analyze the use of the word evolution in Darwin's seminal book.

3. Computer-Assisted Reading and Analysis of Text (CARAT)

Developed at first for numerical and mathematical data processing, computer technology however had its biggest impact on the processing of a wide range of other types of symbols, mainly those of natural language. This shift in technological development and priorities, while stimulating the development of new applications in the domain of text reading and analysis, also confronted researchers in this field to new difficulties, specifically related to textual data. One of the most important and obvious relates to the nature of text itself: even if texts are sequences of natural language symbols, made of letters, words and sentences, their organization

don't qualify as a standard linguistic object and thus cannot be dealt with using standard grammatical, linguistic or logical tools for natural language analysis and processing (Mann and Thompson, 1988; Meunier et al., 2005: 959). In fact, text reading and analysis even fail to qualify as algorithmic or computational processes: a text not only can't be read or analyzed by simply following the natural and linear course of a text, sentence after sentence, but it also never reveals itself fully, its comprehension and meaning being always context- knowledge- and culture-relative (Halliday and Hasan, 1976; Grawitz, 1996; Sprenger-Charolles, 1997; Meunier, 1997; Meunier et al., 2005; Viprey, 2006).

«Each reader and each expert has his or her own repertoire of techniques, procedures, and best practices, which underlie his or her personal and subjective interpretation. […] This highly complex interpretive process cannot be translated into easy and clear-cut algorithms» (Meunier et al., 2005).

Due to these data-processing limitations, the best computers can do in their present technological state is structure data in such as way as to emphasize textual regularities facilitating the interpretation task or make explicit patterns that would be difficult to identify otherwise (Meunier et al., 2005: 974). This particular task is the main objective of Computer-Assisted Reading and Analysis of Text (CARAT).

Resulting from over 50 years of various contributions in philosophy, psychology, sociology, literature, textual semiotics, political science, and history, CARAT aims at offering to professional and expert text readers "assistance in attaining various aspects of the informational or semiotic content of a text" through the development and use of mathematical and algorithmic textual data-processing techniques and tools specifically intended for the construction of the user's own interpretive paths" (Meunier et al., 2005: 961). The *Laboratoire d'Analyse Cognitive de l'Information* (LANCI) specializes in the application of these automated strategies to the reading and analysis of philosophical texts (Chartier et al., 2008; Danis, 2009; De Pasquale and Meunier, 2003; Forest and Meunier, 2000; Meunier, 1997; Meunier et al., 2005; Meunier et al., 1999). This approach has shown good results so far and compares very positively with more linguistically oriented techniques; its greatest advantage is the saving in processing time and its usefulness in processing large textual corpora (Meunier et al. 1999; Meunier et al., 2005; Meunier et al., 2006).

However fruitful this approach may have been in the past, the study of the concept of 'evolution' in the *Origins of Species* and the challenges arising from the concept's scarce lexicalization in Darwin's seminal book call for a new methodology. The present project aims at furthering this original approach by "digging deeper" into the conceptual structures hidden in textual data by designing an iterative clustering method.

4. Methodology

Theoretically speaking, the new text mining algorithm sketched here rests on two fundamental assumptions: 1) the inferential nature and dimension of a concept are linguistically expressed in a differentiated, contextualized and regularized manner; 2) these regularities and patterns can be identified or distinguished using algorithmic, iterative and automatic clustering methods.

4.1. Matrix constitution

Before such processing can be done, however, the whole *Origin of species* must be converted into a matrix, following these steps: each of the 9.442 different words of the *Origin of species* becomes a basic information unit; each of the 974 paragraphs of the work becomes a vector

of the matrix; each paragraph-vector of the matrix has 9.442 dimensions, corresponding to the different words-units in the *Origin of species*; the value of a paragraph-vector dimension is determined by the frequency of the corresponding word-unit in the corresponding paragraph. In the end, Darwin's seminal book is converted into a matrix of 974 by 9442, comprising a total of as much as 9.196.508 numbers!

4.2. Initial matrix clustering

Once this is done, an initial clustering of the matrix is made. The clustering algorithm chosen for this task as well as for the whole conceptual analysis process is K-Means. The standard K-means procedure (Tab. 2) consists in having 1 mean and thus 1 cluster for each 10 segments; in the present case, the 974 vectors constituting the *Origin of Species* matrix will be grouped in 98 clusters on a lexical similarity basis, measured in terms of vector distance.

1- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids,

2- Assign each object to the group that has the closest centroid.

3- When all objects have been assigned, recalculate the positions of the K centroids.

4- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Source: Matteo Matteucci's Tutorial on Clustering Algorithms, Politecnico di Milano: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

Table 2: K-means Clustering Algorithm (MacQueen, 1967)

4.3. Iterative concordance clustering

Following this initial clustering, the iterative concordance clustering process starts by retaining only the most characteristic word of each cluster containing either *evolution*, *evolved* and *evolve*, that is, the word that has the highest TF.IDF (Term Frequency – Inverted Document Frequency) rating each of these clusters. Since there are only 13 paragraphs in which these words occur, this first step will reveal a maximum of 13 clusters and thus 13 characteristic words. Then, the concordance of each of these characteristic words is extracted by regrouping all the paragraphs of the *Origin of Species* in which the said word occurs, and the same process of clustering, cluster selection, TF.IDF rating and ranking, word selection and concordance extraction is performed on each of those new concordances, until no new characteristic word is found or no more clusters containing *evolution*, *evolved* or *evolve* are found. The different steps of this algorithm are shown in Tab. 3.

1. Concordance extraction:	For each cluster containing the word(s) to be analyzed, extract the concordance of the highest-TF.IDF-ranked word.
2. Concordance clustering:	For each previously unselected word, proceed to the clustering of its concordance.
3. Iteration:	Return to step 1, unless 1) no new highest-TF.IDF-ranked word is found, or 2) no clusters containing the word(s) to be analyzed are found.

Table 3: Iterative Concordance Clustering Algorithm

The development of this text processing application is still at an early stage; however, some early conclusions might be drawn from the results obtained.

5. Results and interpretation

In order to identify the principal lexical constituents of the concept of evolution and determine whether or not this underlying conceptual structure includes references to both embryological and specific processes, two different extraction procedures were made: the first one only aimed at the word "evolution", while the second one also added "evolve" and "evolved". The graph in Fig. 1 shows the results of the first analysis.

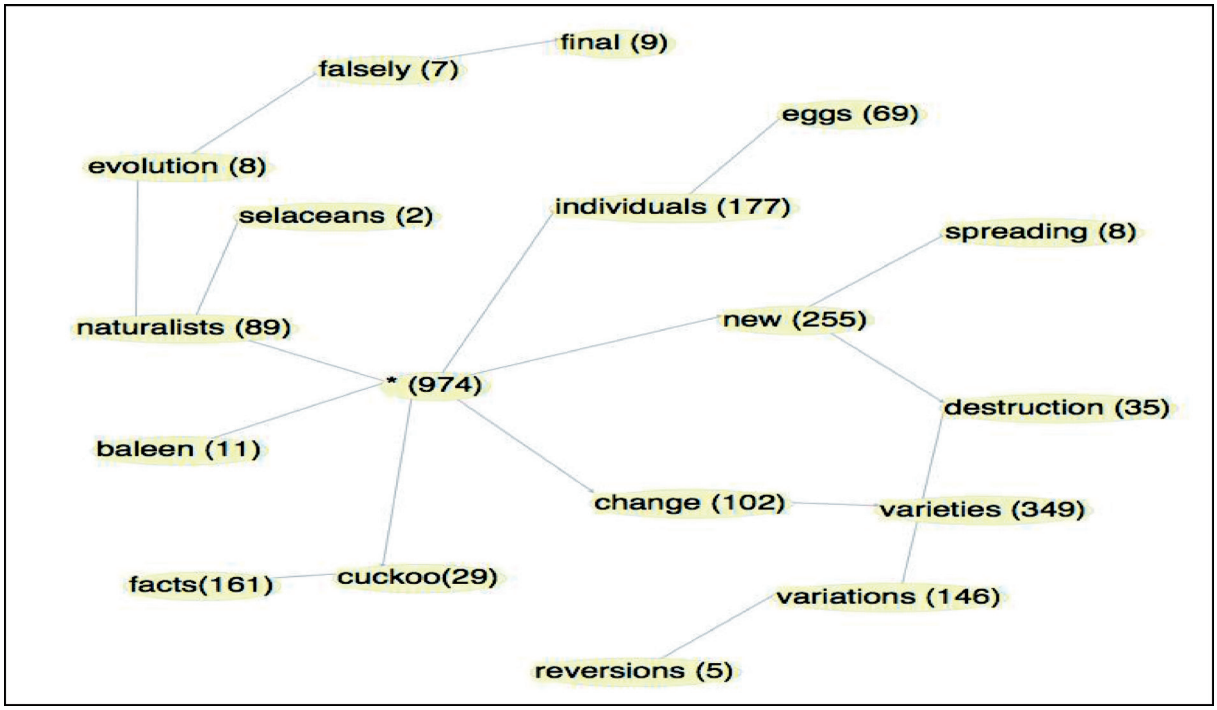

Figure 1: Network visualization of the conceptual analysis of "evolution"

5.1. Conceptual analysis of "evolution"

The central point of the network shows the 974 paragraphs of the *Origin of Species*. After the initial clustering of the whole corpus, the 8 paragraphs in which "evolution" occurs were clustered into six different classes, thus giving the iterative function six initial clustering paths. The highest-TF.IDF-ranked words for each of these classes form the innermost circle of the graph. Of these, only one stops short: it is the one with the word "baleen" as its most characteristic word. Of the rest, four ("naturalists", "change", "cuckoo", "new") dig into what seems to be the specific dimension of the word "evolution": in addition to a couple of names referring to groups of different taxonomic rank (cuckoos, selaceans), words like "varieties", "variations", "spreading", "reversions" and "change" are all closer to specific aspects of development than of embryological ones. As for the last starting path of analysis, "individuals", the embryological dimension seems pregnant, especially as the clustering of this word's concordance leads to "eggs" as being the most characteristic of the class in which "evolution" occurs. However, the chain stops there, for want of any co-occurrence of "eggs" and "evolution", as if Darwin specifically intended not to associate "evolution" with "eggs", which doubtlessly constitutes one of the words with the strongest embryological connotation.

Graph in Fig. 2 shows that the addition of "evolve" and "evolve" in the second extraction procedure radically changed the analysis' results.

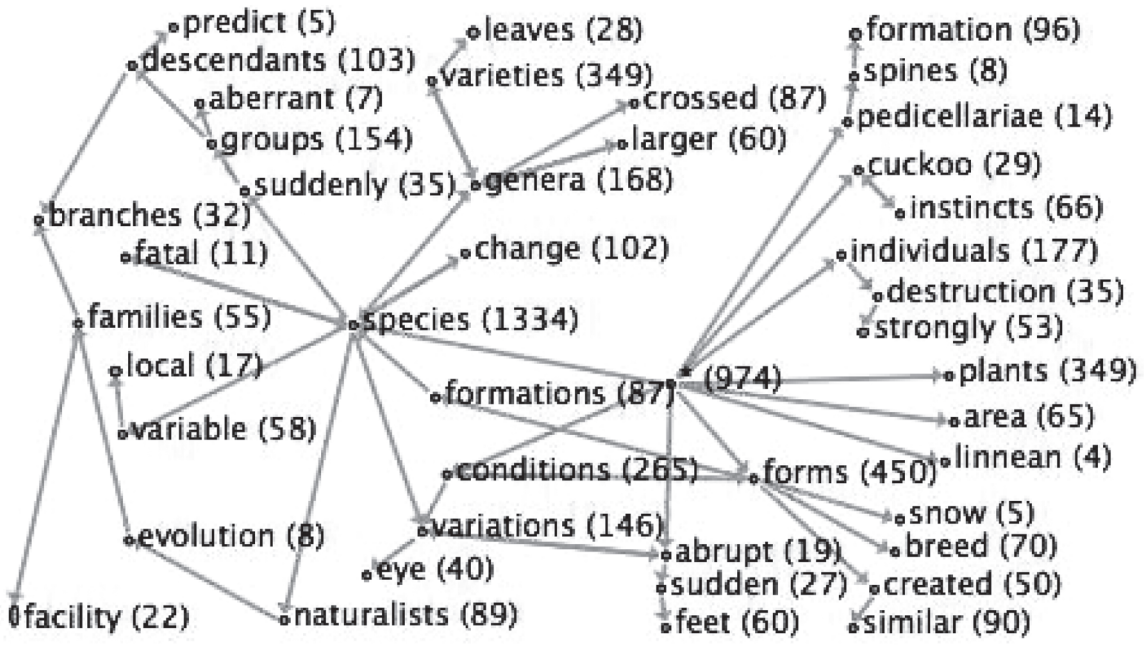

Figure 2 : Network visualization of the conceptual analysis of "evolution", "evolve", and "evolved"

5.2. Conceptual analysis of "evolution", "evolve" and "evolved"

At first glance, what surprises most about this second analysis is the number of words obtained, the algorithm extracting nearly twice as many words as the earlier one. Incidentally, some of these new extracted words play a crucial role in Darwin's transmutation theory: species, forms, genera, conditions, formations, variable, groups, families, branches, descendants, instincts. What is also very interesting about the second analysis is that references are made to words that have already been processed: "formations" and "conditions" both refer to "forms", "variations" to "abrupt", "species" and "conditions" to "variations". At first, these inferential "redundancies" were considered and sought as eventual stopping conditions for the iterative process; however, in light of the results obtained, it now seems as though additional relevance or significance is given to the words refered to more than once. The same thing can also be said of the many words that refer to each other: "cuckoo" and "instincts", "species" and "genera", "species" and "fatal", "species" and "change", "species" and "formations", "variations" and "abrupt", "genera" and "larger", "genera" and "varieties". Indeed, these "bidirectional connections" do more than simply stop the extraction process in a particular ramification of the conceptual network: they also seem to indicate a strong semantic connection between the two words. In short, these inferential "pseudo-redundancies" and "cross-references" both contribute to enrich the conceptual network obtained in the present research by "weighting" the words and the connections it contains respectively. More importantly, in the perspective that these inferences, jointly with the conceptual network as a whole, are more related to specific than embryological development, it seems reasonable to conclude from this second analysis that the richer, more complex conceptual network it produces, besides offering a more exhaustive outline of the concept of evolution, also tends to confirm (or at least doesn't challenge) the conclusion of the first analysis relatively to Darwin's stance on the dual use of "evolution".

Altogether, results and interpretations from both conceptual analyses seem to bring the sixth edition of the *Origin of Species* closer to the contemporary works of the more mature Herbert Spencer, who began to deemphasize the connection between embryology and the

general process of "evolution" and thus contributed to forge the present, strictly specific and most commonly known biological use of the word "evolution". In his *Principles of Biology* (issued 1863-1864), Spencer specifically entitled the chapter on embryology "Development" and gave a footnote explaining that «"development" and "evolution" refer to different types of processes […] Transmutation had now become the chief biological aspect of the system of universal evolution […], and it is probable that Spencer's position on this point marks the beginning of a decline in the […] use of the term in an embryological context» (Bowler, 1975: 108).

Of course, these interpretations, along with the results that made them possible, and the method that originated all this, are not in any way definitive. Further improvements and modifications in the iterative concordance clustering process are to be expected, which will probably alter the results obtained as well as their interpretation. For instance, no stemming was performed on the *Origin of species* prior to clustering; such an operation will however be performed at a later developmental stage, and it is expected that it will lead to the emergence of new "characteristic words". Furthermore, instead of only choosing the highest TF.IDF-ranked words of each cluster in which "evol-" occurs, more precise and sophisticated term selection methods might be chosen, a decision which could also significantly modify the conceptual analysis' results. However important these eventual improvements may be, it is reasonable to expect that the theoretical outcome of the present research will still prevail: the concept of evolution in the *Origin of species* seems more akin to its modern, "spencerian" and solely specific account than its earlier, both embryological and specific, use.

6. Conclusion

The completion of this computer-assisted analysis of the concept of evolution in the *Origin of Species* might prove useful on various grounds. Theoretically speaking, the research should help emphasize the distinction between the semantic associations that constitute specific concepts and the latter's embodiment in natural language, as well while as the possibility of "reading between the lines", that is, of identifying concepts on the sole basis of the lexical associations that constitute them. As for textual data, while bringing new insights in the understanding of the concept of evolution, the present project might also shed new light onto the conceptual analysis of the Darwinian theory itself and on digital philology, hermeneutics and text interpretation in general.

On a more technological basis, the completion of the present project should also help to give "a more rigorous idea of what reading and analyzing a "text" with the help of a computer involves" (Meunier et al., 2005). While such computer-assisted techniques "can diminish the burden for many researchers in the field of social sciences and humanities" (Meunier et al., 2005), they haven't yet made their mark in the humanities and social sciences. The completion of this research and the diffusion of its results precisely aims to help these scientific communities recognize the full importance of computer-assisted text applications and technology.

References

Bowler P.J. (1975). The Changing Meaning of 'Evolution'. *Journal of the History of Ideas*, 36: 95-114.

Bowler P.J. (1984). *Evolution: the History of an Idea*. Berkley: University of California Press.

Chartier J.-F., Meunier J.G. and Jendoubi M. (2008). Le travail conceptuel collectif: une analyse assistée par ordinateur du concept d'accommodement raisonnable dans les journaux québécois. In Heiden, S. et Pincemin, B., editors, *JADT 2008*, Lyon, 12-14 mars, pp. 297-307.

Danis J. (2009). Le concept d'«évolution» dans l'œuvre de Bergson: une analyse conceptuelle appliquée à un corpus philosophique. *Cahiers du LANCI*, (2009-01).

Darwin Charles. (2002-2009). *The Complete Works of Charles Darwin Online*. http://darwin-online.org.uk.

De Pasquale J.-F. and Meunier J.G. (2003). Categorization techniques in computer-assisted reading and analysis of text (CARAT) in the humanities. *Computer and the Humanities* 37 (1): 111-118.

Firth J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Philological Society, Special Issue.

Forest D. and Meunier J.G. (2000). La classification mathématique des textes: un outil d'assistance à la lecture et à l'analyse de textes philosophiques. In Rajman, M. and Chappelier, J.-C., editors, *Proceedings of JADT 2000*, Lausanne, pp. 325-329.

Gravitz M. (1996). *Méthodes des sciences sociales*. Paris: Dalloz.

Halliday M.A.K and Hasan R. (1976). *Cohesion in English*. London: Longman

Harris Z. (1954). Distributional structure. *Word*, 10 (23): 146-162.

Harris Z. (1991). *A Theory of Language and Information: A mathematical approach*. Oxford: Clarendon Press.

Huxley T.H. (1894). Evolution in Biology. In *Collected Essays, vol II: Darwiniana*. London: Macmillan.

MacQueen J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, pp. 281-297.

Mann W.C. and Thompson S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, (8) 3: 243-281

Meunier J.-G. (1997). La Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) comme système de traitement d'information. *Sciences Cognitives*, 22: 211-223.

Meunier J.G., Forest D. and Biskri I. (2005). Classification and Categorization in Computer-Assisted Reading and Text Analysis. In Cohen, H. and Lefebvre, C., editors, *Handbook of Categorization in Cognitive Science*, New York: Elsevier, pp. 955-978.

Meunier J.-G., Remaki L. and Forest D. (1999). Use of classifiers in computer-assisted reading and analysis of text (CARAT). In *Proceedings of CISST'99*, Las Vegas, Nevada, U.S.A., pp. 437-443.

Richards R.J. (1992). *The Meaning of Evolution: the Morphological Construction and Ideological Reconstruction of Darwin's Theory*. Chicago: University of Chicago Press.

Sprenger-Charolles L. (1997). Sciences cognitives et acquisition de la lecture. *Revue Française de Linguistique Appliquée*, 2 (2): 35-49.

Viprey J.-M. (2006). Structure non-séquentielle des textes. *Langages*, 40 (163): 71-85.