Reading Darwin Between the Lines: A Computer-Assisted Analysis of the Concept of Evolution in *The Origin of Species*

Sainte-Marie, Maxime B.
msaintemarie@gmail.com
Université du Québec à Montréal, Canada

Meunier, Jean-Guy
meunier.jean-guy@uqam.ca
Université du Québec à Montréal, Canada

Payette, Nicolas
nicolaspayette@gmail.com
Université du Québec à Montréal, Canada

Chartier, Jean-François
chartier.jf@gmail.com
Université du Québec à Montréal, Canada

Whereas Darwin is nowadays considered the founder of the modern theory of evolution, he wasn't the first to use this word in a biological context: indeed, the word "evolution" already had two distinct biological uses at the time the *Origin of Species* was first published (Bowler, 2003; Huxley, 1897): "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, 1998: 4).

Deriving from the Latin *evolutio*, which refers to the scroll-like act of unfolding or unrolling, the word «evolution» was first used in biology to refer to the development of the embryo, mainly through the formulation, promulgation, and justification of preformationist and epigenetical theories. Embryological evolution would receive its fullest, most modern experimental and theoretical account in the works of Karl Ernst von Baer: characterizing embryological development as a gradual differentiation process leading from homogeneous matter to the production of heterogeneity and complexity of structure, von Baer would usually use the word *Entwickelung* to refer to this dynamic phenomenon, often followed by the Latin *evolutio* in parentheses. The ground-breaking importance of von Baer's work, as well as its diffusion in the scientific community through numerous translations, commentaries, and appropriations, significantly contributed to consecrate the embryological use of the word "evolution".

As for the use of the word evolution to describe specific development, its emergence is closely tied to Lamarckism: even though Lamarck never used the word 'evolution' himself to refer to the transformation of species over time and generations, his commentators, detractors, readers and followers often did however, thus contributing to the semantic alteration of the term. Indeed, "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): 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" (Richards, 1992: 69); von Baer, in rejecting the possibility of transmutation and the popular idea that embryological development recapitulates the progression of the species, used the word «evolution» to refer to both processes; in England, naturalists such as Charles Lyell, Joseph Henry Green, Robert Grant and Richard Owen also used the word "evolution" to both comment and reject Lamarckism (Bowler, 2003; Richards, 1993).

While this dual usage of the word and its most common synonyms at the time (transformation, development, transmutation…) has been confirmed in the works of the most important biologists and naturalists of the first half of the 19th century, little is known about Darwin's own stance on this matter: did he or not use the word 'evolution' or any other word to refer both to embryological and specific development? This question, however crucial it may appear, proves very difficult to answer: while the *Origin of Species* 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 itself is rarely used by Darwin, the

sole and slight exception being the sixth and last edition (1872) of the work.

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*: XV (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)

Occurrences of "evolution", "evolve", and "evolved" in the Origin of Species

This lexical scarcity doesn't necessarily mean however that the concept of evolution isn't present elsewhere in the text, where the words 'evolution', 'evolved', and 'evolve' don't appear. According to distributional semantics theory, meaning can be more easily stated as a property of word combinations than of words *per se*: in every sentence and paragraph, each word brings its own constraints to the whole, reduces the sets of possible words that could fit with it, therefore increasing the total information conveyed and structuring the semantic dimension of each word thus combined. In short, this theory holds 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). In this sense, if concepts are thought of as networks of such meaning-bearing word combinations, then, conceptual structures can determine the semantic dimension of a text without being properly lexicalized; in other words, such considerations, while emphasizing the distinction between the semantic associations of specific concepts and their embodiment in natural language, also seem 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 designation.

In view of this, the fact that the word "evolution" itself is rarely found in the sixth edition of the *Origin of Species* doesn't necessarily imply that the lexical and inferential network 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 of 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 of the same period. However, dealing with word combinations manually might prove difficult, if not impossible. In light of this, a new computer-assisted conceptual analysis tool has been developed by the LANCI laboratory, one which aims to "read Darwin between the lines", that is to identify where the author "conceptually" refers to evolution, regardless of the presence or the absence of the word itself.

Theoretically speaking, this new approach is based 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. Concretely, the algorithm aims at "digging deeper into data" by means of an iterative clustering process. Following an initial clustering of the analyzed corpus (in this case, the 974 paragraphs of the sixth edition of the *Origin of Species*), the iterative concordance clustering process starts by retrieving the most characteristic word of each cluster containing the word(s) to be analyzed, that is, the word that has the higest TF.IDF rating (Term Frequency – Inverted Document Frequency) for each of these clusters. Then, the concordance of each of these characteristic words is extracted from the corpus, 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 the word(s) to be analyzed are found.

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.

Iterative Concordance Clustering Algorithm

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". Figures 1 and 2 show the results of the two analyses.

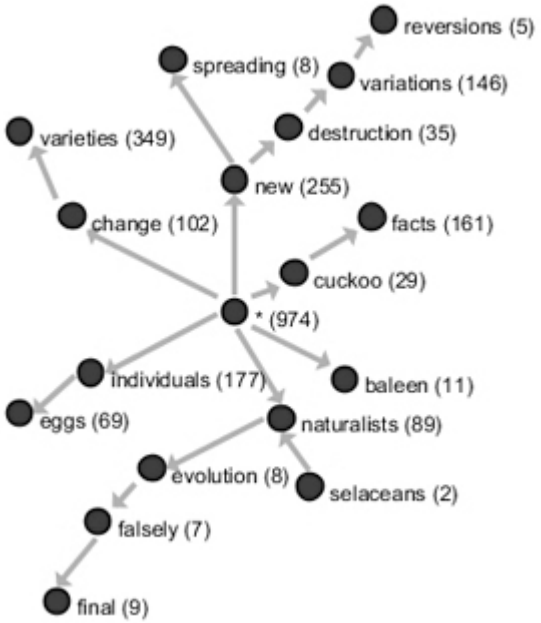

Figure 1: Conceptual analysis of "evolution"

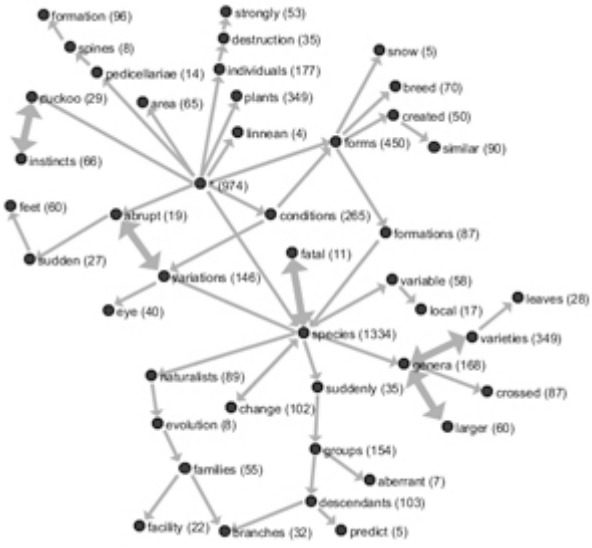

Figure 2: Conceptual analysis of "evolution", "evolve", and "evolved"

In addition to new and unforeseen methodological discoveries, interpretation of both conceptual analyses seems to bring the sixth edition of the *Origin of Species* closer to the contemporary works of the more mature Herbert Spencer, who began to de-emphasize 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".

These results, along with the method that made them possible, are not in any way definitive, and further improvements and modifications of the iterative concordance clustering process are to be expected. Upon completion, this rather new and original approach, while hoping to bring new insights in the understanding of the *Origin of Species*, also aims at underlining the pertinence and usefulness of text mining methods and applications for expert and specialized text reading and analysis, as well as their importance for the future development of philology, hermeneutics, social sciences and humanities in general.

References

Bowler, P.J. (2003). *Evolution: the History of an Idea*. Berkeley: University of California Press.

The Complete Works of Charles Darwin Online. http://darwin-online.org.uk (accessed 25 February 2010).

Harris, Z. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

Huxley, T.H. (1894). 'Evolution in Biology'. *Collected Essays, vol II: Darwiniana*. London: Macmillan, pp. 187-226.

MacQueen, J. B. (1967). 'Some Methods for classification and Analysis of Multivariate Observations'. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp. 281–297. http://projecteuclid.org/euclid.bsmsp/1200512992.

Meunier, J.G., Forest D. and Biskri, I. (2005). 'Classification and Categorization in Computer-Assisted Reading and Text Analysis'. *Handbook of Categorization in Cognitive*

Science. Cohen, H. and Lefebvre, C. (ed.). The Hague: Elsevier, pp. 955-978.

Network Workbench Tool. http://nwb.slis.indiana.edu.

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