The concept of evolution in the *Origin of Species*: a computer-assisted analysis

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

Université du Québec à Montréal (UQAM), Laboratoire d'analyse cognitive de l'information (LANCI), Montreal, Quebec, Canada

Abstract

At the time Darwin first published the *Origin of Species*, the word 'evolution' was used by most biologists of the time to refer not only to specific development, as is the case today, but also to embryological development. Darwin's own stance in that matter is however open to debate, his rare use of the word making it hard to determine whether it is strictly specific or dual, and thus whether the author's conception of evolution is representative or ahead of its time. While this situation certainly stimulates philological, historical, and philosophical debates, it however complicates any attempt to settle the matter on a strict lexical basis, thus making standard text-mining techniques ineffective. To address this specific issue, a computer-assisted method for 'reading Darwin between the lines' is here attempted and described: by using an iterative concordance clustering algorithm, this approach aims at 'digging' into Darwin's concept of evolution as found in the sixth edition of the *Origin of Species*, regardless of any proper designation. In light of the results thus obtained, the concept of evolution in the sixth edition of the *Origin of Species* appears closer to its modern and strictly specific interpretation, inferences made to words related to embryological development being rather rare.

Correspondence:
Maxime B. Sainte-Marie,
Université du Québec à Montréal (UQAM),
Laboratoire d'analyse cognitive de l'information (LANCI),
455 René-Lévesque E., Suite W-5350, Montreal, Quebec, Canada H2L 4Y2
E-mail:
msaintemarie@gmail.com

Whereas Darwin is nowadays considered the founder of the modern theory of evolution, he was not the first to use this 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' had two distinct biological uses at the time the *Origin of Species* was first published: first, to refer to embryological development; and second, 'to characterize the general belief that species have descended from one another over time' (Richards, 1992).

The word 'evolution' was first used in biology to refer to the development of the embryo, mainly through the formulation, promulgation, and justification of both preformationist and epigenetical theories. Through the diffusion, adoption (by Étienne Renaud Serres, notably), and rejection (by von Baer, among others) of Lamarck's ideas of transmutation and of recapitulationism (the idea that embryological 'evolution' recapitulates specific 'evolution'), the word evolution later came to be also related to specific development: '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). This dual use of the word was

Literary and Linguistic Computing, Vol. 26, No. 3, 2011. © The Author 2011. Published by Oxford University Press on behalf of ALLC. All rights reserved. For Permissions, please email: journals.permissions@oup.com

doi:10.1093/llc/fqr019 Advance Access published on 16 May 2011

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021

still well in effect at the time Darwin wrote his perennial book: in his 1852 essay 'The development hypothesis', Herbert Spencer (1820–1903) 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).

1 Embryological and Specific Evolution in the *Origin of Species*: Lexical and Conceptual Issues

Despite the well-established nature of this semantic duality, Darwin's own stance on this specific issue in the *Origin of Species* is rather hard to determine: while the *Origin of Species* is generally considered as the birth document of the theory of evolution, studies on and around this book often emphasize the fact that the word itself as well as its derivatives are rarely used by Darwin. In the first (24 November 1859), second (7 January 1960), third (March 1861) and fourth (June 1866) editions, there is only one occurrence related to 'evolution: evolved' is the last word of the last sentence of the work, the famous 'Tangled Banks' passage. 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. Only in the last edition (1872) are the term 'evolution' and its derivatives more extensively and systematically employed (for a more detailed account, see Table 1) (Darwin, 2002–09).

This lexical scarcity has often been stressed by scholars and specialists. Most of them consider this paucity intentional, as thought Darwin expressly wanted to avoid the word: 'for Darwin, and many other English naturalists of his era, ''evolution'' was an ambiguous word. It evoked various theories that closely associated the history of species with both individual development and an overall progressionist interpretation of the history of nature' (Gayon, 2003).

The fact that the word itself is rarely found does not necessarily mean, however, that Darwin's concept of evolution is without substance: even though 'evolution', 'evolved', and 'evolve' do not appear often, the concept they refer to may well lay elsewhere, 'between the lines'. In order to understand this and make the following conceptual analysis possible, a fundamental cognitive and functional distinction must thus be made between the processes of conceptualization and lexicalization. For that matter, a quick glimpse into distributional semantics (Harris, 1991) might be in order.

Table 1 Occurrences of 'evolution', 'evolve', and 'evolved' in the *Origin of Species*

First edition (one)	evolved: XV (490)
Second edition (one)	evolved: XV (490)
Third edition (one)	evolved: XV (525)
Fourth edition (one)	evolved: XV (577)
Fifth edition (two)	evolved: XIV (573), XV (579)
Sixth edition (fourteen)	evolution: (VII: 201(two), 202), VIII (215), X (282), XV (424 (three))
	evolve: VII (191)
	evolved: VII (191, 202(two)), XV (425, 429)

According to this theory, meaning can be more easily stated as a property of word combinations than of words *per se*: 'similarities and patternings among the co-occurrence likelihoods of various words correlate with similarities and patternings in their types of meaning' (Harris, 1991). In every sentence and paragraph, each word brings its own constraints to the whole, reduces the sets of possible words that could fit therein, therefore increasing the total information conveyed and structuring the semantic dimension of each word thus combined. Thus, meaning is not 'in' the words themselves but 'between' them, that is, in the word combination networks in which they take part, that determine what other words might co-occur with them and what overall meaning might thus be conveyed. Now, since the linguistic expression of a concept is structured and constrained by these word combinations, it would then be possible to analyze concepts on the basis of the lexical co-occurrence networks that regularly express them, whether these concepts are properly lexicalized or not.

In view of this, the fact that the word 'evolution' itself is rarely found in the sixth edition of the

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021

Origin of Species does not necessarily imply that the lexical environments in which it appears and that shape its meaning cannot be found elsewhere in the text (i.e. where the concept is not properly lexicalized) in a rather similar fashion (except for the occurrence of the word 'evolution', of course) and cannot be analyzed in order to better extract the lexical co-occurrence network shaping the concept's meaning. In other words, taking into account text segments similar to those where the word 'evolution' occurs 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. For that purpose, the use of text-mining strategies, which help to emphasize textual regularities and patterns that would be difficult to identify otherwise (Meunier *et al.*, 2005), might seem here natural, even necessary. However, due to the lexical scarcity of 'evolution', 'evolve', and 'evolved' in the *Origin of Species*, a new text-mining methodology is needed, one that aims to identify the word combinations shaping up the semantic dimension of concepts, independently of their lexicalization.

2 Methodology

Theoretically speaking, the concept-mining approach developed and described here is based on a few linguistic and cognitive assumptions, shown in Table 2.

The main objective of the following method is to extract the text segments most similar to those in which the word referring to the analyzed concept occurs, in order to identify as many of the different, concept- and semantic-shaping, lexical co-occurrence and inference patterns as possible.

In order to accomplish this, the following method has been developed: after an initial clustering of the corpus to be analyzed, the word that has the highest TF.IDF rating (Term Frequency – Inverted Document Frequency) is extracted for each cluster in which a text segment containing the word analyzed (here 'evolution', 'evolve', or 'evolved') is found. This should allow for the extraction of the words that have the strongest co-occurrence link with the term to be analyzed.

Then, a concordance for each of these retrieved words is created by collecting each text segment in which the said word occurs, and the same operations of clustering, cluster selection, TF.IDF rating and ranking, occurrence-based word selection, and concordance extraction are performed on each of those new concordances, until no new highest TF.IDF-ranked word is found or no more text segments containing the word(s) analyzed are found in the clusters. Table 3 summarizes the different steps of this concordance-clustering algorithm:

This approach is not entirely automatic, though, given that some intervention is required for preprocessing the corpora to be analyzed, choosing the clustering algorithm and parameters as well as interpreting the results obtained. However, since these specific activities are rather characteristic of any computer-assisted text analysis method, the present algorithm might be considered as automatic as it could possibly be.

Table 2 Fundamental hypotheses of the present concept-mining method

1	A concept is linguistically expressed in a differentiated, contextualized, and regularized manner by a lexical network that shapes its semantic dimension through co-occurrence constraints.
2	Meaning conferred by lexical co-occurrences is independent of the corresponding concept's lexicalization.
3	The meaning of a concept can be analyzed through lexical co-occurrence patterns that are similar to those in which the given concept is properly lexicalized.
4	The lexical co-occurrence network specific to a given concept can be identified through concordance analysis.
5	The lexical co-occurrence network characteristic of concepts can be formally represented using vectorial algebra ('bag of word' method), given that the corpus to be analyzed is converted into a matrix.
6	Lexical co-occurrence pattern identification analysis can be done using algorithmic and automatic clustering methods

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021

Table 3 Iterative concordance clustering algorithm

1. Concordance extraction	For each cluster containing the word(s) 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.

3 Preliminary Experimentations and Results

In order to apply this method to the concept of evolution in the *Origin of Species*, a few decisions in parameterization had to be taken beforehand. Since the sixth edition of the *Origin of Species* is the one in which 'evolution' and its derivatives occur the most, it has been chosen as the experimentation corpus for the present research. The preprocessing was done as follows: each of the 9442 different words that the sixth edition contains is considered a basic information unit, and each of its 974 paragraphs a vector of the whole matrix. Each vector of this matrix has thus 9442 dimensions, corresponding to the different information units contained in the corpus; the value of each vector dimension is then determined by the frequency of the corresponding unit in the corresponding paragraph.

Clustering was done using *K*-Means, at a ratio of one centroïd for each ten segments. Also, 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 analyses were made: the first one only aimed at the word 'evolution', while the second one also added 'evolve' and 'evolved'. Results of these two analyses are shown in Figs 1 and 2 respectively.

Figure 1 shows the results of the first analysis, that of 'evolved'. Of all the conceptual paths that were 'dug', only one stops rather shortly: 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 such as '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.

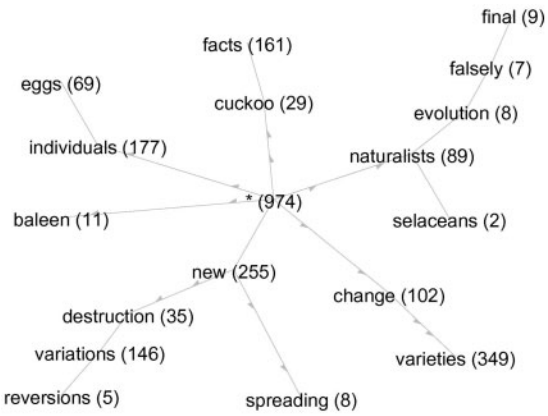

Fig. 1 Conceptual analysis of *evolution*

Figure 2 shows that the addition of 'evolve' and 'evolved' radically changes the analysis' results. Some of the new words extracted play a crucial

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021

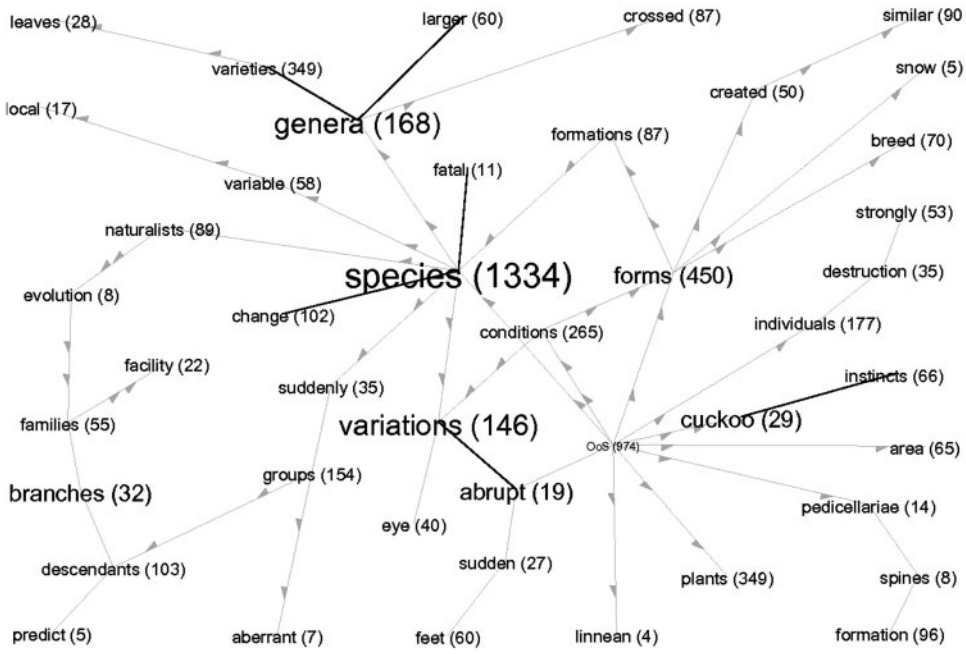

Fig. 2 Conceptual analysis of *evolution*, *evolve*, and *evolved*

role in Darwin's transmutation theory: species, forms, genera, conditions, formations, variable, groups, families, branches, descendants, and instincts. References are made to words previously processed, and such inferential 'redundancies' seem to give them additional relevance of significance: 'formations' and 'conditions' both refer to 'forms'; 'variations' to 'abrupt'; 'species' and 'conditions' to 'variations'; 'formations', 'fatal', 'formation', 'change', and 'genera' to 'species'; 'species', 'larger', and 'varieties' to 'genera'. Many words also refer to each other, and such bidirectionality seems to be the sign of a strong semantic connection between the words implied : 'cuckoo' and 'instincts', 'species' and 'genera', 'species' and 'fatal', 'species' and 'change', 'species' and 'formations', 'variations' and 'abrupt', 'genera' and 'larger', and 'genera' and 'varieties'. Since these inferential 'pseudo-redundancies' and 'cross-references', jointly with the conceptual network as a whole, tend to confirm the conclusion of the first analysis relatively to Darwin's stance on the dual use of 'evolution', it seems reasonable to conclude that the concept of evolution in the *Origin of Species* refers mostly to specific development, inferences made to embryological development being rather rare.

4 Interpretations and Further Improvements

The development of this concept-mining procedure is still at an early stage; however, some early conclusion might already be drawn. Indeed, results from both analyses tend to bring the sixth edition of the *Origin of Species* closer to the contemporary works of the more mature Herbert Spencer, who began in the 1860s 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'. Such a conceptual shift is obvious in his *Principles of Biology* (1863–1864): in this work, the chapter on embryology is specifically entitled 'Development' and a footnote is added by the author, 'explaining that

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021

''development'' and ''evolution'' refer to different types of processes' (Bowler, 1975). The present research seems to show that Darwin's concept of evolution in the sixth edition of the *Origin of Species* follows, in a rather implicit or indirect way, Spencer's own treatment of the concept in its lexicalized form.

Of course, these interpretations, along with the results and method that made them possible, 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* in the processing stage; such an operation might allow for the emergence of new 'characteristic words'. It is also expected that different clustering algorithms may produce different results, as a previous study from the laboratory has shown (Chartier *et al.*, 2010). Furthermore, instead of only choosing the highest TF.IDF-ranked words of each cluster in which 'evolution', 'evolve', or 'evolved' 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.

References

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

Chartier, J.-F., Meunier, J.-G., and et Djellali, C. (2010). Analyse des variations entre partitions générées par différentes techniques de classification automatique de textes. In Bolasco, S., Chiari, I., and Giulano, L. (eds), *Proceedings of the 10th Internationnal Conference Journées d'analyse statistique des Données Textuelles*. Rome: Edizioni Universitarie di Lettere Economica Diritto, pp. 37–48.

Darwin, C. (2002–09). *The Complete Works of Charles Darwin Online*. http://darwin-online.org.uk (accessed 28 February 2011).

Gayon, G. (2003). From Darwin to today in evolutionary biology. In Hodge, J. and Radick, G. (eds), *The Cambridge Companion to Darwin*. Cambrige: Cambridge University Press, pp. 240–66.

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

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. (eds), *Handbook of Categorization in Cognitive Science*. The Hague: Elsevier, pp. 955–78.

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

Downloaded from https://academic.oup.com/dsh/article/26/3/329/1143129 by guest on 22 September 2021