Editorial genesis

From comparing texts (product) to interpreting rewritings (process)

Rudolf Mahrer | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Rossana De Angelis | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Andrea Del Lungo | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Almuth Grésillon | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Jean-Louis Lebrave | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Valentine Nicollier Saraillon | "Manuscript – Linguistics – Cognition" team ITEM (ENS
| CNRS) LATTICE (CNRS
| ENS
| Université Sorbonne nouvelle Paris 3)

Thierry Poibeau | "Manuscript – Linguistics – Cognition" team ITEM (ENS

Editorial genesis

From comparing texts (product) to interpreting rewritings (process)

Rudolf Mahrer, Rossana De Angelis, Andrea Del Lungo, Almuth Grésillon, Jean-Louis Lebrave, Valentine Nicollier Saraillon, Thierry Poibeau, Frédérique Mélanie-Becquet & Bénédicte Vauthier
"Manuscript – Linguistics – Cognition" team
ITEM (ENS/CNRS)
LATTICE (CNRS/ENS/Université Sorbonne nouvelle Paris 3)

> Nothing is ever definitively said while we still have time ahead of us, held out to the speaker as a promise.
>
> (*Rien n'est jamais définitivement dit tant que le temps est devant soi, donné comme espérance au locuteur.* Peytard 1993, §121)

In literary genetics, "editorial genetics" deals with the "public life" of texts, whereas the writing process is affected by edition and diffusion. Editorial genetics frequently has to deal with cases of "editorial rewriting": in the literary domain for example, authors frequently modify previously published works, so that several versions may co-exist. We are especially interested in Balzac's *La Bourse* (translated in English as *The Purse*) since we know three authorized versions of this specific work.

By comparing different texts associated with a single work, the literary geneticist is facing different products that are themselves the result of a writing process. However, different specificities should be outlined: (1) the writing process does not leave any trace: we just have access to different products/texts and (2) since the texts we compare seem to be achieved, differences must be referred, not to programmatic or temporary linguistic structures, but to the reconfiguration of a pre-existing textuality.

Do such products still reflect the processes that have given birth to them? Does the comparison between two texts considered as variations of a same text give access to this transformation's processes? After describing the objects of this particular textual comparison and the terminology that permits to give an account of such phenomenon, this contribution suggests to express these

DOI 10.1075/z.194.08mah
© 2015 John Benjamins Publishing Company

questions differently, as a matter of *poetics of transitions between texts*, or, further digging, an *hermeneutics of the transition between texts*.

Keywords: editorial genetics; textuality; variation

1. Genetic criticism and editorial genesis

Genetic criticism emerged in France at the end of the 1960s. Rather than describing written signs and their use, it focuses on their production. In other words, instead of looking at writing as a product, it treats it as a process. This theoretical approach consists in gathering together documents and observations, in order to identify the author's writing habits, linguistic reformulations, and cognitive mechanisms. This offers us an insight into the creative process and allows us to accurately characterize the preparatory documents that constitute the so-called *avant-texte* (also known as the pre- or fore-text).

As text composition is a process that takes place over time, it can be divided into several different *phases*. These are defined not only by the technological devices that are used, but also by the writing purpose and the properties of the successive preparatory activities.[1] According to Lebrave (2009)'s schematization, for example (see Figure 1), the initial stage of *traceless gestation* is followed by the first written drafts. Then comes the so-called *accommodations* phase paving the way for publication. The fourth phase consists of *revisions* to the published text. The fifth and final phase (*alterations*), concerns unauthorized modifications, in particular those made after the author's death. In this phase, which lies beyond the scope of *genetic criticism*, the text's story continues, with new writers and therefore new geneses.

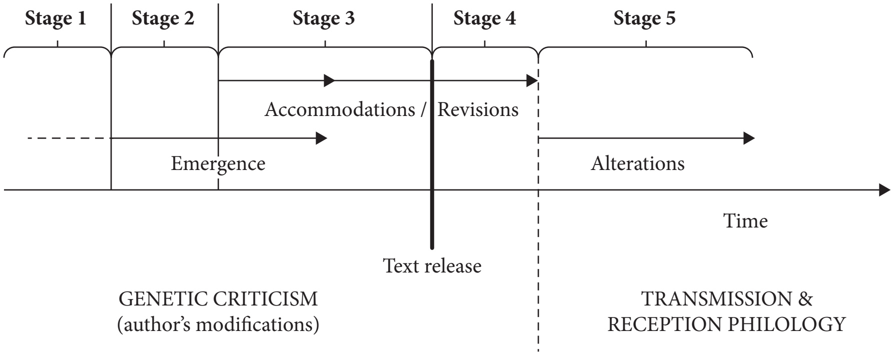


Figure 1. Lebrave's schematization (2009, 18)

[1] According to our chosen theoretical perspective, writers do not just prepare a written sign (product) open to interpretation. They prepare a trace, but also – *in fine*, a reading (process).

On first examination, *editorial genesis* deals with the fourth stage in Lebrave's schema, wherein "textual changes [are] made by the author after the text's publication" (Lebrave 2009, 18). More specifically, it concerns the point when a written text emerges from the writer's private sphere and embarks on the route to publication. In the draft phase, the writer (or writing authority) is – in principle – the sole agent and guarantor of all aspects of text production, but in this new phase, these processes are ordinarily carried out by a plethora of actors, including typists, correctors, proofreaders, printsetters, master printers and publishers.

2. Rethinking the frontier between product and process, text and *avant-texte*

By viewing writing as a process rather than a product, genetic criticism has encouraged research on preparatory documents (in particular manuscripts) bearing visible signs of text generation. As we have seen, geneticists use the term *avant-texte*, or even *dossier génétique* (Grésillon 1994, 109), to refer to the set of (more or less) chronologically ordered preparatory documents for a given text. In this opposition between *avant-texte* and text, the former is associated with the process and the latter with the product, but how does one move from the one to the other?

2.1 *Avant-texte* and text

Characterized by its original approach to literary work, textual genetics traditionally views the pass for press as a cut-off point. When the writer signs this pass, he or she agrees to the publication of the text as it stands. This contract marks the frontier between the third and fourth stages in Lebrave's schema (see Figure 1).

> Thus, the preparatory stage, where everything remains possible, gradually moves into a *new dimension* where the author's intervention is (with a few *exceptions*) far more limited. (De Biasi 2005 [2000], 45, our italics.)

> If we compare the physical, finished text with the virtual and unfinished manuscripts from a genetic point of view, we can say that *its emergence coincides with its publication* – an operation that turns a private, autographic object into a public, allographic text. (De Biasi 2000, 29, our italics.)

Such a representation of the genetic domain assumes that the reassessment of a published text is very much an exception, and further questioning of the editorial process beyond this cut-off point is discouraged. Nevertheless, authors do intervene after publication as attested by tree different and nonexlusive manifestations (1) variously justified requests by the author for *corrections*, (2) documents found in the author's personal papers containing superficial or in-depth revisions, (3) variations between different editions.

Research undertaken by the Manuscript – Linguistics – Cognition team, which has been studying post-publication rewriting since 2012, has consistently confirmed its prevalence. It concerns all periods and all genres. This observation has a corrosive effect on the genetic field, as both Grésillon (2007; Montaigne's *Essais* (*Essays*), Eluard's *Donner à voir*, Aragon's *Communistes* and Ronsard's complete works), and Vachon (2009; Balzac's work) noted:

> But the *terminus ad quem* [of the *avant-texte*] is equally complex. The criterion that was initially selected for the pass for press would rapidly prove illusory. Two aspects of textual genetics show just how fragile it was: successive editions (reviewed, corrected and augmented by the author), and the production process of theatrical works. In the former, the pass for press would be followed by publication, but the author would frequently annotate his/her personal copy of the work, introducing changes and additions to be incorporated in the next edition. (Grésillon 2007, 32)

> In other words, the Balzacian corpus leads us to formalize a specific theory for post-publication genetics that challenges the opposition between text (the focus of literary criticism) and *avant-texte* (focus of genetic criticism) and even makes it irrelevant, by showing that these are just two sides-one private, the other public-of the same reality. (Vachon 2009, 41)

We could also mention authors such as Erasmus, La Bruyère, Molière, Flaubert, Claudel, Ramuz, Cendrars, Césaire, Duras and Genet, who were all rewriters to some degree. From the *editorial genetics* point of view, we need to reconsider at least two aspects of the connection between text and *avant-texte*, that is, between written documents deemed to contain evidence of processing and documents regarded as products. First of all, we have to agree that, in theory, nothing can bring the process to an end once and for all. Nothing can sever the relationship between a creator and his or her work–apart from the disappearance of one or other. This is the meaning behind Peytard's opening quote which sums up the condition of human creation in the field of linguistic works.

Second, we need to distinguish between two categories of documents within the genesis of a work. Some writings are produced in preparation for the upcoming production of an oral or written discourse. They contain discursive forms of the preparation (preparatory notes, plans, scenarios, drafts, etc.), and use writing tools (manuscripts, pencils, quill pen, typewriters, screen and keyboard, etc.) and techniques (crossing out, erasing, copying, pasting, etc.) that connect the writing space and writing gesture with the content (invention and organization) and verbal expression.[2] Other genetic documents contain the completed form of a given

[2] Mahrer and Nicollier (2015) refer to this document family as *écrits de la préparation du discours*, or simply *écrits de préparation*, and suggest treating it as a type of discourse that encompasses several genres.

discourse genre, and can only be regarded *a posteriori* as part of the *avant-texte*, when we observe that they are not, after all, the final link in a chronological chain of steps ending with the final textual product.

Editorial genetics challenges critics who study writing processes via documents principally illustrating the second case. There is often no way of telling whether or not a text that has been published (once or even several times) will subsequently be revised by its author. The latter may not even have planned to do so him- or herself. Balzac, for example, writing in 1834 about the fourth edition of *La Peau de chagrin* (*The Magic Skin*) (published by Werdet in 1835), stated that "its purpose was to give once and for all a final form to the texts belonging the large general edition of [his] works". However, regarding the fifth edition (by Delloye and Lecou in 1838), he commented that "the text […] has been revised with such care that it must be regarded as the only one that exists, such is the difference between it and previous editions",[3] although it is worth mentioning that Balzac later went on to produce two more versions of that same novel. From the writer's perspective at least, these editorial genetic *steps* are viewed as definitive *stages*:

> A final state can very easily revert to a draft, and what seemed to imply the end of a creative process can evolve into a new writing episode. The text returns to being an *avant-texte*. The pen replaces the lead type. (Grésillon 2007, 32)

2.2 Publication seen in a genetic light

Every writing can be rewritten. This observation leads us to reject the idea that, in principle, publication marks the end of the writing process. This does not mean that the geneticist should regard the published text as just another draft. In accordance with the objective of genetics as a discipline, the study of editorial genesis can be defined as the description of the effects of publishing on the various writing processes and, at the same time, the study of these processes under specific circumstances. As this chapter is too short for us to discuss it in detail, we simply outline its key aspects:

1. At a semiotic level, in the publishing phase, the piece of writing is modified in that its medium and even its semiotics change. The bundles of sheets or handwritten notebooks may become a poster, a leaflet, or a book. This aspect extends from spatial constraints and techniques to the rewriting gesture.
2. At a pragmatic level, the piece of writing is modified after having been put to one side for a time owing to publishing constraints (commercial and/or

[3] *Lettres à l'étrangère*, Paris, Calman Lévy, p. 195–196 & 454–455 (quoted by Falconer 1969, 72).

technological).[4] The persistence of the graphic trace, combined with the work's prestige, allow for its continuing genesis. The longer this period lasts, the more "the I who was previously writing changes", as Grésillon (2007, 32) wrote, thinking not only about Montaigne writing his "*allongeails*" (expansions), but also about himself. The historical context and writing fashions (genre, style, etc.) change as well, giving the writer further reasons for revising his or her text.[5]

3. Writing is modified when it becomes *public*. This adjective has to be understood in two different ways.
   a. On the addresser's side, the enunciatory entity becomes openly collective, as the production of the artefact requires a range of skills (the writer is joined by the copyist, publisher, director of the collection, editor, printer, bookseller, etc.). The picture of the writer preparing his manuscript alone at his desk no longer matches reality. Henceforth, the publication of a text does not simply involve its author, be it on a practical, economic, legal, or even moral level. Genetics should be not only about the history of texts, but also about the history of the work itself. It cannot ignore the collective dimension of the publication, but must describe the product of *editorial enunciation* (Souchier 2007).
   b. The presence of an audience means that there is also an addressee's perspective: once published, a text can prompt questions and elicit a response, be it positive or negative. Rewriting can then be interpreted as a sort of ratification or resistance to the criticisms formulated by the entities to which the work was submitted.
4. Finally, in *literacy*, publication is an important sociocultural act. The author and publisher commit themselves to the release and, more particularly, to the published book itself. Thus, the symbolical value that the author places on publishing (his or her *imagining* of the publication act) will condition his or her post-publication rewriting (from Erasmus, who judged it as necessary and boundless, to Hugo who regarded it as vain and refused to do it).

[4] The contract for the second edition of *La Peau de chagrin* (*The Magin skin*) was signed in August 1831, two weeks after the first edition had sold out. The newer version came out the same year. Balzac made his revisions to the Furne edition sometime between the publication of the 15th volume of the *La Comadie humaine* (*The Human Comedy*) in 1846 and his death in 1850.

[5] Regarding the rewritings and reprints of *La Peau de chagrin* spanning the period from 1831 to 1846, Falconer makes the hypothesis that "a whole side of a 'flamboyant romanticism' from the primary text" that was probably pleasing to the 1831 readership, was gradually abandoned as readers' tastes changed. (Falconer 1969, 73, who mentions the disappearance of "play on words or on typography, exotic names and adjectives, jests, and paradoxes").

In short, from the perspective of research on the writing process, publication raises the status of a text to that of a *written document exerting constraints on its virtual rewritings*. These rewritings have to be interpreted in the light of these constraints.

2.3 What should be the *object* of linguistic analysis?

In order to describe the various writing processes, genetic critics can scrutinize the creation of discourse both on and off line (e.g. using Eye and Pen software). They can also study production by examining the traces left by certain writing technologies (e.g. pen and paper, or keyboard and hard drive) or even accounts of those practical experiences. Evidence of the various phases of editorial genesis (book corrected by its author, poster, correspondence between author and publisher, etc.) includes the different editions available in libraries. Although the latter do not contain autographical traces of rewritings, by comparing the texts, geneticists interested in this aspect can see that there are several versions of the same work.

Comparing these versions brings all kinds of differences to light. In the following section, we show how data processing can facilitate these comparisons, thereby promoting the study of hitherto neglected features. First, however, we need to discuss the terms used to think about and describe these differences. How can they be articulated with the notion of rewriting used in genetics to talk about writing processes?

Philology offers us the notion of *variant*, which is often used in the genetic context, but its philological origin can lead to misinterpretation.[6] The notion of *variation* seems more relevant, even though it is not used in the genetic field. As in music, where it is used in relation to the notion of *themes,* variation presupposes the existence of an identity that serves as its basis. Whereas variants are mutually exclusive, the term variation implies the coexistence of two objects, regarded as different from one angle, but identical from another (Ferrer 2011).

In the language context, the linguistic analysis of variation involves the comparison of two sequences (A and B), extracted from two versions of what is considered being the same text. We can say that there are *linguistic variations* between Version A and Version B if the comparison between the two enables us to pinpoint both identical sequences and, in *n* point(s), different sequences. *Variation* is thus defined as the relationship between Sequence x_i from Version A and Sequence y_i from Version B.

[6] The philological variant presupposes that there is only one authorized and pertinent text, and that this text is an original one. By contrast, for genetic criticism, the finished text does not yet exist and is an end in itself.

Compact basis (b_i) refers to the left and right co-texts in Versions A and B that flank the x_i and y_i sequences.

2.4 From variation to *rewriting*

For geneticists, concerned with the analysis of writing processes, *rewriting* is a written reformulation by which the previous state (x) of a linguistic sequence is invalidated by its replacement with a new one (y). From this point of view, rewriting can be regarded as a substitution process whose elements are chronologically oriented (Lebrave 1983). We can access it in at least three different ways: by observing the gestures involved in the operation, either on or off line (Leblay & Caporossi 2014; see also Leblay and Caporossi this volume); by observing the graphic traces of these gestures (e.g. crossings-out); and by studying the products of the operation. The latter implicitly brings together rewriting and variation. For a variation to be used as a basis for genetic investigation, it needs to correspond to a writing gesture. However, not all variations between texts are the outcome of rewriting. Hence, for a text comparison to take place in a genetic investigation (formulating hypotheses about the writing processes), Versions A and B need to be (1) regarded as the same text, (2) attributable to the same production entity, and (3) chronologically oriented.[7] Ideally, (4) Version B should also result directly from Version A, with no other documents between the two (the precision of genetic hypotheses depends on this constraint). Under these conditions, Versions A and B can be viewed as the input and output of a machine whose internal workings are concealed, rather like black boxes. The geneticist attempts to model the writing processes by comparing Sequences x_i and y_i. This modelling naturally depends first and foremost on the metalanguage used to interpret x_i and y_i, as well as on the *laws* governing their transformation in terms of form and content.

3. Analyzing Balzac's work in the light of editorial genetics

Balzac's work provides an exemplary case study for editorial genetics. Unlike other writers of his time, who mainly revised and rewrote their work while it was still in manuscript form (Flaubert is a prime example), Balzac regularly made major modifications between successive editions of a same text (up to eight different versions of *La Peau de chagrin*, something that is extremely rare in literary practice.

[7] Although the relation of variation is reciprocal (if A is a variation of B, B is a variation of A) the relation of rewriting is not (if B is a rewriting of A, A cannot be a rewriting of B).

Let us summarize the different phases of the genetic process, as critics have already described it.[8] The *dossier génétique* relating to Balzac's works (collected at the end of the 19th century by Viscount Spoelberch de Lovenjoul) contains only a few preliminary documents (notes, drafts, plans, scenarios), but a welter of manuscripts. Their analysis shows that after some hesitations (as revealed by many stalled beginnings), writing seemed to progress quickly, as evidenced by the small number of deletions or insertions (to the extent that one could imagine that these texts are just copies of previous versions that have disappeared). The first rewriting phase often occurred during the correction of the galleys and proofs, in some cases reflecting a desire to expand the text with large, and even spectacular, insertions. The second phase took place after publication, in a highly systematic way (for instance, Balzac started correcting the Furne edition of the *La Comédie humaine* (*The Human Comedy*), first published in 1842, in his personal copy).

Stressing the specificity of this practice, Stéphane Vachon notes that "Balzac's writing process include[ed] the continuous re-publication of his works and recreat[ed] his manuscripts by multiplying the working editions" (Vachon 1997, 72). As we said in the Introduction, we do not consider Balzac's rewriting habits to cancel the difference between the preparatory documents and the published editions, but this difference no longer corresponds to a strict divide between process and product. Moreover, it should be noted that the rewriting process operates differently when it takes place directly in the proofs as opposed to between the different editions of a text: the former often involves an extension of the text, whereas the latter is a kind of perpetual re-adaptation, maintaining a certain compactness of text, especially for the Furne edition of *La Comédie humaine*. With the notable exception of *La Peau de chagrin*, which was augmented with the addition of several narrative episodes, the new version of each text is nearly always slightly shorter than the previous one that served as the basis for the rewriting process.

This is the case of *La Bourse* (*The Purse*), a short story that summarizes the main features of the editorial genetic process in Balzac's work. The original version of this story was published in the second edition of *Scènes de la vie privée* (*Scenes of Private Life*) by Mame et Delaunay-Vallée in 1832. The second version was published in 1835 by Madame Béchet, in Volume IX of *Études des mœurs au XIX^{ème} siècle* (*Studies of Manners in the 19th Century*), then with les *Scènes de la vie parisienne* (*Scenes of Parisian Life*). The third version corrected by the author (the 1839 Charpentier edition of les *Scènes de la vie parisienne* contained the same text as the previous one) is the one that was published in the Furne edition of *La*

[8] See Stéphane Vachon's article entitled "Les enseignements des manuscrits d'Honoré de Balzac. De la variation contre la variante", *Genesis* 11, 1997.

Comédie humaine (Volume I, 1842), so the story returned to its original place in *Scènes de la vie privée*. The fourth version incorporated the handwritten corrections that Balzac had inserted in his personal copy of the Furne edition for a new edition that was only published after his death.

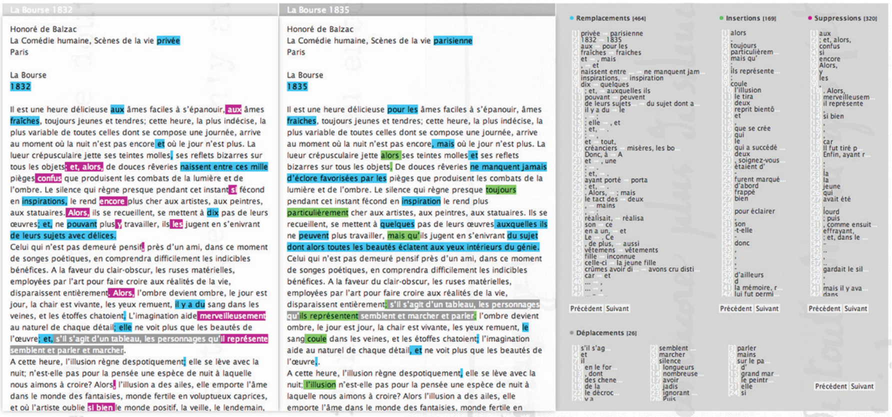
4. The automatic analysis of literary variants

Despite having different goals, researchers belonging to the literary genetics and philological communities all have to deal with different versions of a given text that can vary immensely and thus generally contain a great many variants. The manual analysis of these variants requires a huge amount of work first to list and classify them, then to analyze their changes over time. It is a highly repetitive and tedious task, which leads to a large number of errors when the work is done manually. Computers are thus extremely useful for automatically listing and classifying these variants.

Literary genetics takes its name, of course, from biological genetics. The parallel goes further, as a text can be seen as a sequence of words or even characters, just as a strand of DNA is a chain of four different nucleotides represented by the letters A, T, G and C). Two DNA sequences can be compared on the basis of four formal operations: deletion, insertion, substitution (aka mutation), and frame shift (Lewis 2005). In the same way, two texts can be compared using these four operations: the author may have added some words and removed others, and a portion of text may have been replaced or moved. It should, however, be noted that from a strictly formal point of view, *deletion* and *insertion* can account for every type of change, as a substitution or a move can be formalized as a deletion followed by an insertion. However, the notion of *move* is more informative, as it expresses the fact that the same piece of text has been removed from one place in order to be added in another place. Similarly, *substitution* indicates that one sequence of letters (or words) has been replaced by another at a precise point in the text. Different parameters have to be defined for these operations, such as the minimum length of what can be classified as a move (e.g. changes in punctuation are generally regarded not as moves but as deletions or insertions of commas, full stops, etc.).

A number of tools have been developed to track changes in texts and classify these changes according to the four operations described above. These tools can generally be configured so as to tailor the parameters to the context or the author. One of the best known piece of software is Edite/Medite, developed under the supervision of Ganascia and Lebrave at ITEM (Bourdaillet et al. 2009).

Edite/Medite requires the documents to be formatted in XML. All the changes between two versions of a text can then be automatically calculated and classified in one of the above four categories. A graphical interface has independently been



La Bourse 1832

Honoré de Balzac
La Comédie humaine, Scènes de la vie privée
Paris

La Bourse
1832

Il est une heure délicieuse aux âmes faciles à s'épanouir, aux âmes fraîches, toujours jeunes et tendres; cette heure, la plus indécise, la plus variable de toutes celles dont se compose une journée, arrive au moment où la nuit n'est pas encore et où le jour n'est plus. La lueur crépusculaire jette ses teintes molles, ses reflets bizarres sur tous les objets; et, alors, de douces rêveries naissent entre ces mille pièges confus que produisent les combats de la lumière et de l'ombre. Le silence qui règne presque pendant cet instant si fécond en inspirations, le rend encore plus cher aux artistes, aux peintres, aux statuaires. Alors, ils se recueillent, se mettent à dix pas de leurs œuvres; et, ne pouvant plus y travailler, ils les jugent en s'enivrant de leurs sujets avec délices.
Celui qui n'est pas demeuré pensif, près d'un ami, dans ce moment de songes poétiques, en comprendra difficilement les indicibles bénéfices. A la faveur du clair-obscur, les ruses matérielles, employées par l'art pour faire croire aux réalités de la vie, disparaissent entièrement. Alors, l'ombre devient ombre, le jour est jour, la chair est vivante, les yeux remuent, il y a du sang dans les veines, et les étoffes chatoient. L'imagination aide merveilleusement au naturel de chaque détail; elle ne voit plus que les beautés de l'œuvre; et, s'il s'agit d'un tableau, les personnages qu'il représente semblent et parler et marcher.
A cette heure, l'illusion règne despotiquement; elle se lève avec la nuit; n'est-elle pas pour la pensée une espèce de nuit à laquelle nous aimons à croire? Alors, l'illusion a des ailes, elle emporte l'âme dans le monde des fantaisies, monde fertile en voluptueux caprices, et où l'artiste oublie si bien le monde positif, la veille, le lendemain,

La Bourse 1835

Honoré de Balzac
La Comédie humaine, Scènes de la vie parisienne
Paris

La Bourse
1835

Il est une heure délicieuse pour les âmes faciles à s'épanouir, âmes fraiches, toujours jeunes et tendres; cette heure, la plus indécise, la plus variable de toutes celles dont se compose une journée, arrive au moment où la nuit n'est pas encore, mais où le jour n'est plus. La lueur crépusculaire jette alors ses teintes molles et ses reflets bizarres sur tous les objets. De douces rêveries ne manquent jamais d'éclore favorisées par les pièges que produisent les combats de la lumière et de l'ombre. Le silence qui règne presque toujours pendant cet instant fécond en inspiration le rend plus particulièrement cher aux artistes, aux peintres, aux statuaires. Ils se recueillent, se mettent à quelques pas de leurs œuvres auxquelles ils ne peuvent plus travailler, mais qu'ils jugent en s'enivrant du sujet dont alors toutes les beautés éclatent aux yeux intérieurs du génie.
Celui qui n'est pas demeuré pensif près d'un ami, dans ce moment de songes poétiques, en comprendra difficilement les indicibles bénéfices. A la faveur du clair-obscur, les ruses matérielles, employées par l'art pour faire croire aux réalités de la vie, disparaissent entièrement; s'il s'agit d'un tableau, les personnages qu'ils représentent semblent et marcher et parler; l'ombre devient ombre, le jour est jour, la chair est vivante, les yeux remuent, le sang coule dans les veines, et les étoffes chatoient; l'imagination aide au naturel de chaque détail, et ne voit plus que les beautés de l'œuvre..
A cette heure, l'illusion règne despotiquement, elle se lève avec la nuit; l'illusion n'est-elle pas pour la pensée une espèce de nuit à laquelle nous aimons à croire? Alors l'illusion a des ailes, elle emporte l'âme dans le monde des fantaisies, monde fertile en

● Remplacements [464]

1 privée → parisienne
2 1832 → 1835
3 aux → pour les
4 fraîches → fraiches
5 et → , mais
6 , → et
7 naissent entre … → ne manquent jam…
8 inspirations, → inspiration
9 dix → quelques
10 ; et, → auxquelles ils
11 pouvant → peuvent
12 de leurs sujets… → du sujet dont a…
13 il y a du → le
14 . → ;
15 ; elle → , et
16 ; et, → .
17 ; → ,
18 et → tout,
19 créanciers → misères, les bo…
20 Donc, à → A
21 et → , une
22 ; → .
23 ; et, → .
24 ayant porté → porta
25 ; et, → .
26 . Alors, → ; mais
27 le tact des → deux
28 , → mains
29 , → .
30 réalisait, → réalisa
31 son → ce
32 en a un, → et
33 Le → . Ce
34 , de plus, → aussi
35 vêtemens → vêtements
36 fille → inconnue
37 celle-ci → la jeune fille
38 crûmes avoir di… → avons cru disti…
39 car → et
40 … → ,
41 … → ;
42 … – → .

● Insertions [169]

1 alors
2 .
3 toujours
4 particulièrem…
5 mais qu'
6 .
7 ils représente…
8 ;
9 coule
10 l'illusion
11 le tira
12 deux
13 reprit bientô…
14 et
15 ,
16 que se crée
17 qui
18 le
19 qui a succédé …
20 deux
21 , soignez-vous…
22 étaient d'
23 -
24 furent marqué…
25 d'abord
26 frappé
27 bien
28
29 pour éclairer
30 ,
31 son
32 -t-elle
33 -
34 donc
35 ,
36 ,
37 ,
38 -
39 d'ailleurs
40 d
41 la mémoire, r…
42 lui fut permi…

● Suppressions [320]

1 aux
2 ; et, alors,
3 confus
4 si
5 encore
6 Alors,
7 y
8 les
9 .
10 . Alors,
11 merveilleusem…
12 il représente
13 ,
14 si bien
15 ,
16 ,
17 ,
18 car
19 Il fut tiré p…
20 Enfin, ayant r…
21 ,
22 ;
23 la
24 la
25 jeune
26 qui
27 avait été
28 ,
29 lourd
30 ; puis
31 , comme ensuit…
32 effrayant,
33 ; et, dans le …
34 –
35 ..
36 ,
37 ..
38 gardait le sil…
39 ,
40 ,
41 mais il y ava…
42 dans

● Déplacements [26]

1 s'il s'ag…
2 semblent …
3 parler
4 et
5 marcher
6 mains
7 il
8 silence
9 sur le pa…
10 en le for…
11 longueurs
12 d'
13 , dont
14 nombreuse…
15 grand mar…
16 des chene…
17 avoir
18 le peintr…
19 de la
20 jadis
21 elle
22 le décroc…
23 ignorant …
24 si
25 y a
26 Puis

Figure 2. Screenshot of the graphical interface developed by Yannick Saraillon to complement Edite/Medite software

developed to enable users to navigate within the text, view two versions of the text side by side, and gain access to the complete list of changes corresponding to the four operations, among other things. Figure 2 provides a screenshot of this interface.

Edite/Medite is clearly extremely useful for manipulating different versions of a given text. However, researchers working on such texts, be they geneticists, philologists or linguists, rapidly feel the need for further functionalities. The four operations are purely formal and do not take into account the linguistic content of the sections of text under consideration. A linguistic analysis of these fragments would be highly useful, making it possible to access the changes from a different, more content-oriented, point of view (see Brunner and Pordeus Ribeiro this volume).

The following section contains a proposal to go beyond the current state of the art. Our goal is to define a method for automatically analyzing the variations observed in literary texts. We begin by defining a linguistic model (What kinds of facts do we want to observe? What classification would be useful in our context?). We then examine the extent to which this model can be implemented, taking Balzac's short story *The Purse* as an example.

4.1 The linguistic classification of literary variants

The links between linguistics and literary texts are complex. Whereas linguistics mainly involves the meticulous analysis of a finite set of sentences to test linguistic features on the basis of tiny, systematic variations, establishing *minimal pairs* to check whether these variations are linguistically driven, literary texts come to us *as they are* (Fuchs 1983; Culioli 1983). For example, Culioli reminds us that a literary text is not a representative sample of a linguistic phenomenon, and cannot be extended or directly manipulated, contrary to what linguists frequently do. A text is the result of a complex creative process, but it does not afford us access to the operations that gave rise to the actual observable result.

From this point of view, textual genetics gives us a unique opportunity to gain access to variations, as the different versions of a given text contain traces of the changes that the author (or editor) have made to the text, thereby providing clues to the creation process.

4.1.1 *What kind of linguistics?*

There are obviously many ways of doing linguistics, so some rather naive but nevertheless useful questions are: Which type of linguistic analysis is most useful for the analysis of literary texts? Which linguistic theory is most appropriate? And first of all, what should we expect from a linguistic analysis in this context? To what extent is linguistic different from stylistics? In other words, should we regard stylistics as the branch of linguistic analysis that is best suited to literary texts?

Stylistics is generally assumed to "establish principles capable of explaining particular choices made by individuals and social groups in their use of language" (Wikipedia). From this point of view, stylistics differs from mainstream linguistics. The first task is to observe and describe the differences between two versions of a given text. At this stage, the reasons why the author's choices vary from one version to the other, the interpretation of the text, and the aesthetic dimension of text variation are put to one side.

Our point of departure is thus quite simple. We want to stay as close to the text as possible, which means that the analysis must describe the modifications in the text from a neutral point of view, untrammeled by theory. We stay away from the interpretative dimension (status of the author, stylistic value of any change, etc.) as far as possible. The interpretative analysis takes place in the second stage, based on the descriptive approach proposed here. The advantage of this is that the stylistic analysis is based on comprehensive observations, and not on isolated examples, as is all too often the case.

We therefore need to develop a general descriptive framework for the linguistic analysis that corresponds to our specific area of interest (literary variants from a genetic analysis perspective) as well as to the more general domain of literary text analysis. We propose dividing this analysis into four different *levels* (words, morphology and syntax, sentences and whole text).

1. Lexicon (richness and diversity; contexts of use)
2. Grammatical morphemes, more specifically:
    a. Determiners
    b. Tense and mood of the verbs
    c. Modals

3. Sentences (length and complexity)
4. Text (structure; organization and use of discourse markers)

These levels are, of course, quite generic and rather conventional. However, they ensure that the analysis is comprehensive and does not only take one aspect of the text into account, which is what often happens.

The next step consists in identifying existing tools and/or developing new ones to automate the analysis.

4.2 Automating the analysis

Lexical analysis requires a system that is capable of compiling lists of words and comparing them. This analysis can be run on either the word forms themselves or a lemmatized version of the text (in order to differentiate between types and tokens). More sophisticated analyses can be conducted using advanced tools like the Sketch Engine, which provides a detailed analysis of the context in which words appear in a text (Kilgarriff et al. 2004).

4.2.1 *Lexical analysis*

We propose to calculate the specificity of each word form, and rank words according to their positive or negative specificity. *Specificity* corresponds to the traditional definition put forward by Lafon (1980).

4.2.2 *Morphological and syntactic analysis*

Morphosyntactic analysis requires a system that can analyze a text and add morphological features to word forms. The quality of analyzers is generally satisfactory, but performances on literary texts can vary dramatically, depending on the nature of the text. Manual correction of the analyzer's output is required to achieve a near-perfect analysis.

Syntactic analysis requires a syntactic parser, but parsing is far from perfect, and manual correction can be overwhelming. However, in the case of variants between two versions of a text, sometimes only a local syntactic analysis is required, rather than the analysis of whole sentences.

4.2.3 *Implementation*

So far, our strategy has consisted in using existing tools as far as possible, rather than developing new ones. Of course, when nothing is available to perform the analysis, the development of new tools becomes necessary. The preliminary experiments described in this chapter all involved existing, off-the-shelf tools.

The lexical analysis was performed by the TXM toolbox (developed as part of the Textometrie project: http://textometrie.ens-lyon.fr/) (Heiden 2010). This tool allows users to examine the vocabulary of a given text and to compare the word lists of different texts, as and when required (see previous section).

As for the morphological analysis, we used TreeTagger (Schmid 1994), which is also integrated into the TXM platform to compare lemmatized vocabulary lists. TreeTagger can further be used to compare the morphosyntactic features of two versions of the same text. For example, we can check how verb tenses, modals or determiners are used, as these are known to be frequent sources of change between different versions. Other features can be studied separately, as long as they form part of the morphosyntactic analysis that is automatically performed. It should be noted that if TreeTagger was used in this study, this was mainly on account of its ease of use, as an integral part of TXM. We have since developed our own analyzer, which perform better than TreeTagger for French, and we will be using this new analyzer in subsequent studies.

Concerning parsing, we did not use an actual syntactic analyzer, as we decided that the sequences of tags produced by TreeTagger represented the first step towards syntax. This proved to be sufficient in this context, but a real parser might be useful in other studies. However, parsers are far from perfect, and this should be set against the cost of manual error correction.

4.3 Experiment

In this section, we describe an experiment featuring a short story by Balzac.

4.3.1 *The corpus*

Instead of studying the genesis of a text from the early drafts to first published version, we chose to focus on the notion of *editorial genesis* which, as we have seen, refers to the study of the different published versions of a text. We felt that this was more relevant in our case, as we were not so much interested in the creative process itself as in the practicalities of our analytical method. For example, Balzac is known for having published several versions of most of his texts (there are at least eight extant versions of *La Peau de chagrin*, some featuring major changes, others just a handful of minor corrections).

We analyzed the first two versions of the short story *La Bourse* (*The Purse*), published in 1832 and 1835, although there are also two later versions, published in 1839 and 1842. A truly comprehensive analysis would, of course, take all four versions of the text into account.

With this type of text, geneticists and/or linguists look for different and even contrasting things. What major changes did the text undergo? Are there any regularities/patterns in the changes Balzac made to the text? By the same token, are there any isolated, remarkable changes? This, in our opinion, is what makes the analysis of literary texts especially difficult: analysts want to have access both to the generalities and to the specificities of the text. Analysis tools should allow both.

4.3.2 *Semi-automatic analysis*

TXM automatically generates vocabulary lists, and offers different ways of comparing these lists, based on frequency and specificity (see Figure 3).

A quick look at the results leads to some interesting observations:

1. Systematic changes, which are easy to spot even without a tool of this sort. For example, the name of the main protagonist was *Jules* in 1832 and *Hippolyte* in 1835;
2. Less systematic changes, which are probably meaningful but difficult to spot when they are distributed across the text. For example, Balzac frequently changed the adverbs he used (*chèrement* and *alternativement* are replaced with *particulièrement* and *complaisamment*). This kind of observation is especially hard to interpret, and requires careful study of the text to see whether it is meaningful or not;
3. Nonsystematic changes owing to orthographic variations. The spelling of some French words was still quite fluid at the beginning of the 19th century, as we can see here, where *mouvemens* becomes *mouvements* and *vêtemens* becomes *vêtements*, while *savant* becomes *savans* and *sergent* becomes *sergens*;
4. Some isolated words or semantic families also undergo changes from one edition to the other (e.g. between the 1832 and 1835 editions, *créancier* disappears and *rançon* appears). These changes can be spotted automatically, but their interpretation requires expert analysis.

1832 version	**1835 version**
Jules	Hippolyte
mouvemens	mouvements
vêtemens	vêtements
savant	savans
sergent	sergens
…	…
créanciers	rançon
…	…
chèrement	particulièrement
alternativement	complaisamment
…	…

Figure 3. Word-form lists

These changes, ranked by specificity (Lafon 1980; Lebart et al. 1997) as well as by frequency, are highly informative. When specificity is taken into account, it is not the absolute number of changes that counts, but this number in relation to the frequency of the word (i.e. a relatively small number of changes can be meaningful if they concern an infrequent lexical form). In TXM, punctuation marks can be regarded as lexical items. The comparison of the two versions of Balzac's text ranked by specificity yields two very interesting results (see Figure 4).

Token	Freq	1832	1835	Spec.
Jules	80	80	0	–23.7
…	93	87	6	–18.7
!	101	67	34	–3.0
:	233	141	92	–2.8
A	7	7	0	–2.1
Alors	44	30	14	–1.9
.	953	515	438	–1.8
et	634	340	294	–1.2
car	27	18	9	–1.2
y	66	39	27	–1.0
…	…	…	…	…
Hippolyte	51	0	51	15.7

Figure 4. List of the most specific changes, from one version to the other

1. Balzac made numerous and meaningful changes to the punctuation. Strong punctuation marks (especially exclamation marks, suspension marks, semicolons and, to a lesser extent, full stops) are less common in the 1835 version of the short story, meaning that this version contains fewer sentences. As a direct consequence, sentences are longer and more complex in the 1835 version. This aspect of the work has never been directly addressed by critics, as far as we know.
2. Discourse connectives are also less frequent in the 1835 version, where Balzac made more use of juxtaposition, placing sentences side by side without any explicit connections between them.

Observations concerning isolated lexical items, as well as punctuation marks, can therefore be meaningful at the sentence level, as we discuss in greater detail below.

The morphological analysis performed with TreeTagger failed to reveal any meaningful changes at this level. The use of modals, verb tenses and determiners seemed to remain relatively stable from one version to the other (which is not the case for all Balzac's novels). We think it is also important to spot instances of stability, as the absence of change can be just as meaningful when interpreting changes in the text.

As for the sentence and discourse levels, we have already observed that sentences are longer and more complex in the 1835 version. The linguistic tricks used by Balzac to make the sentences longer include greater use of personal pronouns and relative clauses.

Automatic analysis can highlight various systematic patterns in discourse particle use. For example, in the 1832 edition, Balzac frequently used a semicolon followed by the French conjunction *car* (for), which is not “stylistically correct” (esp. according to critics of the time). All occurrences of this construction had disappeared in the 1835 version: Balzac either deleted the word *car*, so as to obtain two separate sentences, or else removed the semicolon, leaving the conjunction on its own.

(1) *Cette mission lui plut ; car l’amour fait son profit de tout, et rien ne séduit plus un jeune homme que de jouer le rôle d’un bon génie, auprès d’une femme.* (1832)

Cette mission lui plut. L’amour fait son profit de tout, et rien ne séduit plus un jeune homme que de jouer le rôle d’un bon génie auprès d’une femme. (1835)

The same can be observed for the conjunction *donc* (therefore). A brief look at word forms shows that *donc* remained stable, whereas *Donc* (with the uppercase D) disappeared entirely, meaning that in 1835, it was no longer used to start sentences. From a stylistic point of view, it is supposed to be better to integrate conjunctions within the sentence, rather than just putting them at the beginning.

(2) *Schinner avait choisi ses amis parmi les hommes les plus honorables et les plus distingués.* (1832)

Schinner avait donc choisi ses amis parmi les hommes les plus honorables et les plus distingués. (1835)

This kind of change is probably a consequence of comments Balzac received in the press: he was often portrayed as a writer with a poor style, using colloquial and improper expressions.

5. Conclusion: Which process gives access to a genetic and linguistic analysis of variation?

Genetics shares the variation problem with other textual approaches. However, unlike other points of view that can lead to two texts being treated as variations of the same text (considering every form of discursive re-elaboration, including translation, plagiarism, imitation, vulgarization, transposition, etc.), the genetic outlook on textual variation (a) explores the modalities of passing from Versions A to B, and (b) does so by adopting the point of view of the text’s reader-writer.

In this chapter, we have looked at the conditions and constraints that influence this rewriting activity, particularly when it takes place after the text's publication, thus generating several versions of the *same* text.

Seeking to understand a process, even though it only can observe different products, the genetic approach can be defined as a *poetics of transition between states* (Lebrave 2009), where linguistic analysis is applied to *n* discrete textual variations in order to reconstruct the process whereby Text A *morphs* into Text B.[9] However, the reconstructed process may not literally correspond to the transformation of A into B. In reality, the observed variations merely provide material for a *differential semantics* that consists in comparing the effects of varying sequences on their supposedly identical co-texts. The reconstructed process concerns the motives that lead the reader-writer to prefer the second sequence to the first, and to move from the one to the other. Variation linguistics is not intended to provide writing research with a readymade instrument for analyzing transformation. Rather, it is about the *poetics* or, more accurately, the *hermeneutics* of the *transition between states*. It describes the precise nature of the transition from one state to another in linguistic terms, and this description enables us to formulate hypotheses about the reasons behind this change by the writer, be it consciously or unconsciously, in his or her particular context, depending on how far we can reconstruct his or her linguistic skills.

Our contribution is thus to move from written linguistics to writing linguistics, from the linguistics of product to the linguistics of process.

References

de Biasi, Pierre-Marc. 2005 [2000]. *La Génétique des textes*. Paris: Nathan.

Bourdaillet, Julien, Jean-Gabriel Ganascia, and Jean-Louis Lebrave. 2009. "Topologie et génétique textuelles: un dialogue médié par la machine." *Lexicometrica* 9 special issue: unnumbured (available online: http://lexicometrics.univ-paris3.fr/numspéciaux/special9/bourdaillet.pdf).

[9] If we examine the traces left *by* the writing (and not *in* the writing itself), and if we consider them from a linguistic point of view, the "deepness of relationship between all forms of textual variation" (Lebrave 2009, 18) allows to assert that the *linguistic treatment* of variation and the observable traces in the manuscripts are identical. One should either way seek to understand the differences between what we call *textual variations*.

Culioli, Antoine. 1983. "Préface." In *La genèse du texte: les modèles linguistiques*, ed. by Jean-Louis Lebrave, Catherine Fuchs, Almuth Grésillon, Jean Peytard, and Josette Rey-Debove, 9–12. Paris: CNRS Editions.

Falconer Graham. 1969. "Le travail du style dans les révisions de *La Peau de chagrin*." In *L'Année balzacienne*: 71–106.

Ferrer, Daniel. 2011. *Logiques du brouillon. Modèles pour une critique génétique*. Paris: Seuil.

Fuchs, Catherine. 1983. "Eléments pour une approche énonciative de la paraphrase dans les brouillons de manuscrits." In *La genèse du texte: les modèles linguistiques*, ed. by Jean-Louis Lebrave, Catherine Fuchs, Almuth Grésillon, Jean Peytard, and Josette Rey-Debove, 73–102. Paris: CNRS Éditions.

Grésillon, Almuth. 1994. *Éléments de critique génétique. Lire les manuscrits modernes*. Paris: PUF. DOI: 10.1515/arbi.2000.18.1.1

Grésillon, Almuth. 2007. "'Nous avançons toujours sur des sables mouvants.' Espaces et frontières de la critique génétique." In *La création en acte. Devenir de la critique génétique*, ed. by Paul Gifford, and Marion Schmid, 29–40. Amsterdam/New York: Rodopi.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (PACLIC 24), ed. by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Unemoto, Kei Yoshimoto, and Yasunari Harada, 389–398. Sendai: Tohoku University.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "The Sketch Engine." In *Proceedings of EURALEX*, ed. by Geoffrey Williams, and Sandra Vessier, 105–116. Lorient: Université de Bretagne Sud.

Lafon, Pierre. 1980. "Sur la variabilité de la fréquence des formes dans un corpus." *Mots* 1: 27–165. DOI: 10.3406/mots.1980.1008

Lebart, Ludovic, André Salem, and Lisette Berry. 1997. *Exploring Textual Data*. Dordrecht: Kluwer Academic Press. DOI: 10.1007/978-94-017-1525-6

Leblay, Christophe, and Gilles Caporossi. 2014. *Temps de l'écriture. Enregistrements et représentations*. Paris: L'Harmattan.

Lebrave, Jean-Louis. 1983. "Lecture et analyse du brouillon." *Langages* 69: 11–23. DOI: 10.3406/lgge.1983.1139

Lebrave, Jean-Louis. 2009. "Manuscrits de travail et linguistique de la production écrite." *Modèles linguistiques* 30: 13–21. DOI: 10.4000/ml.330

Lewis, Ricki. 2005. *Human Genetics: Concepts and Applications*. Boston, MA: McGraw Hill.

Mahrer, Rudolf, and Valentine Nicollier Saraillon. 2015. "Les brouillons font-ils texte? Le cas des plans pré-rédactionnels de C. F. Ramuz." In *Faire texte. Unité(s) et (dis)continuité*, ed. by Jean-Michel Adam, (to be published). Besançon: Presses Universitaires de Franche-Comté.

Peytard, Jean. 1993. "D'une sémiotique de l'altération." *Configurations discursives, Semen* 8. Retrieved on 8 April 2014. URL: http://semen.revues.org/4182

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing*, ed. by Harold Somers, and Danny Jones, 44–49. Manchester: NeMLap.

Souchier, Emmanuël. 2007. "Formes et pouvoirs de l'énonciation éditoriale." *Communication et langage* 154: 23–38. DOI: 10.3406/colan.2007.4688

Vachon, Stéphane. 1997. “Les enseignements des manuscrits d’Honoré de Balzac. De la variation contre la variante.” *Genesis* 11: 61–80. DOI: 10.3917/balz.010.0199

Vachon, Stéphane. 2009. “Perspectives en génétique balzacienne.” In *Balzac, Flaubert. La genèse de l’œuvre et la question de l’interprétation*, [Proceedings of the Second International Conference, 14–16 December 2007, Nagoya], ed. by Kazuhiro Matsuzawa, 35–45. Nagoya, Japan: Graduate School of Letters, Nagoya University.