Analysis of non directive interviews with the " bundle " method

Mathieu Brugidou, Pierre Le Quéau

Analysis of non directive interviews with the « bundle » method

by

Mathieu Brugidou
GRETS-EDF
Direction des etudes et recherches
1 avenue du general de Gaulle - 92141 - Clamart
01 47 65 40 53
mathieu.brugidou@edf.fr

Pierre Le Queau
Université Pierre Mendès France- CREDOC
142, rue du Chevaleret, 75013 Paris
01 40 77 85 51
Pierre.Le-queau@upmf-grenoble.fr

Summary: This article presents the results of work on the analysis of non-directive research interviews (ENDR). The identification of a series of terms which are repeated to a significant extent at the same time in a short sequence of text (i.e. 'bundles') enables us to see the 'meaning packets' which make up the episodes. By following these bundles of terms, we hope to be able to locate thematic ruptures (the end of a series of bundles), i.e., the passages where the thread of the narrative unfolds as a new theme (or in this case, an episode) is developed (new series of bundles). The algorithm of the bundles has been incorporated in the semantic analysis software TROPES. We intend to demonstrate the utility of this method for the analysis of non-directive interviews. In particular we will seek to combine a lexicometric approach (bundles formed from graphic forms) and a semantic approach (bundles formed from classes of terms considered as equivalent in terms of meaning).
NDRI, narrative, structural analysis of narrative, analysis of discourse, lexicometry, bundle.

Semi-directive or non-directive interviews are currently employed by sociologists. However, it would seem that the status given to "people's words" is not altogether clear. Either we consider that the interviews constitute a generally reliable source of information (once the information has been extracted from the subjective 'chaff' of the interview) or we see them as a discourse whose singularity is irreducible and which can only be reported word for word, thereby abandoning all hope of real analysis.
While the latter view leaves the interviews intact (with analysis being abandoned for a simple functional expression of the interviews), the former view often leads to a "deconstruction" of the interview. The analysis of the thematic content is the most common methodological approach. The discourse is broken down and reduced to a certain number of themes which are then transformed into hypotheses subjected to quantitative validation.
The uncertainty surrounding the status of the interview is compounded therefore by doubts regarding the analytic methods.
These difficulties appear to be linked, in part, to the fact that we do not consider the interview as a narrative, i.e., a story with a beginning, a middle and an end, whose coherence is ensured by a "plot."
We thus lack knowledge on the social effects specific to the narrative where the narrator puts forward for the interviewer - and sometimes for himself - his own story. We do not fully understand the narrator's operational modes: the identification and narration of individualized moments and their articulation in the form of "episodes" in a coherent "plot" which renders them comprehensible to us.

Indeed, what we must be able to identify is, on the one hand, the different episodes or sequences and, on the other hand, the cohesion of the narrative (which renders intelligible what would otherwise be a series of unlinked episodes). The cohesion of meaning cannot therefore be the proposition but rather the sequence. We propose here a method for highlighting these sequences and the way they are linked together.

The identification of a series of terms repeated at the same time in the same period in a short sequence of text (i.e. 'bundles') enables us to see the 'meaning packets' which make up the episodes. By following these bundles of terms, we hope to be able to locate thematic ruptures (the end of a series of bundles), i.e., the passages where the thread of the narrative unfolds as a new theme (or in this case, an episode) is developed (new series of bundles). The algorithm of the bundles has been incorporated in the semantic analysis software TROPES. We intend to demonstrate the utility of this method for the analysis of non-directive interviews. In particular we will seek to combine a lexicometric approach (bundles formed from graphs) and a semantic approach (bundles formed from classes of terms considered as equivalent in terms of meaning).

1. "Bundles"

1. 1. The dynamics of the narrative

Repetition is the motor of the spoken narrative. The notion of "bundles" reflects this. Research on this form was directly inspired by P. Lafon who highlighted the value of a sequential approach using lexicometry (Lafon, 1981, 1983). By adopting this approach, he demonstrated that it was possible to estimate the regularity of the distribution of a form in a "homogenous text with a beginning and an end". Irregularity seems to be the rule insofar as the forms tend to arrive “in bundles” which are grouped together in a limited part of the text, whereas regularity is rarer and characterized by "athematic" forms. Thus it should be possible to identify the themes and thematic sequences within a spoken narrative.

Within the framework of an analysis of television interviews with politicians[1], a specific method of lexicometric processing was developed which enables us to identify the repetitions which are an inevitable feature of politicians' speech. The program is only interested in a certain number of these repetitions. It identifies and measures the "remarkable" redundancy of certain terms, essentially nouns, which organize the development of the discourse. Within this context, "remarkable" is defined as an irregular distribution of these words.
The method was subsequently tested for the processing of non-directive interviews (of biographical narratives) as part of university research (M. Brugidou & P. Le Queau, 1995). Within this context, the repetition takes on a wholly different meaning. The initial project aimed to identify the succession of themes, without thematic injunction on the part of the interviewer. Although it may appear trivial at first sight, it is interesting to note that in comparison with a television interview with a politician, the narrative of an "ordinary" person talking about himself, appeared infinitely "poorer". The different bundles are much less numerous, but longer. In other words, given comparable "texts", an ordinary person telling his life story uses a far more limited vocabulary than a politician taking part in a television program.

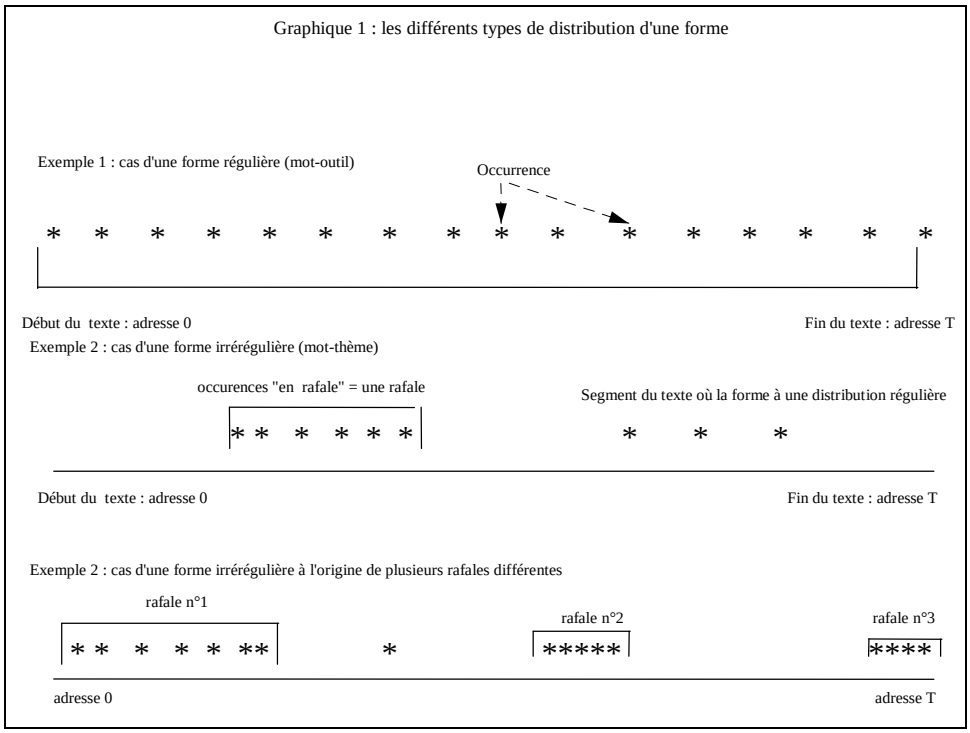
R. Ghiglione's team is responsible for the development and improvement of the processing of bundles in the TROPES software. Discussing psycho-socio-cognitive theory in a recent work, R. Ghiglione developed the idea that the distribution of propositions in a text is subject to certain rules which condition the text's internal coherence. We find in particular the human memory's capacity to "process" a text in only very short cycles: *"five to ten propositions are processed in a cycle, some are kept in a ‘buffer' memory, allowing a before/after link"* (1998, p. 25). Short-term memory (MCT) and its role in cognition is also discussed in *Cognitivo-discursive analysis*, a work which goes into more detail on the production of referential coherence. To sum up, it would seem that repetition, in either literal or anaphoric form, is a way of compensating for memory gaps, and is an essential tool in the production of the internal and sequential coherence of the narrative.

1. 2. The algorithm of bundles

The aim of P. Lafon's work was to characterize the entire distribution of a form. Some of these forms, as we have seen, are known to be very irregular. Others are more frequent. Here we are concerned with identifying all the passages in the text where any form is repeated to a noticeable extent, i.e., overused given its overall frequency in the text. In other words, we are not interested in the overall distribution but rather its localized characteristics. It is in our interest therefore to 'optimize' the bundle, ie., to retain only the most dense part/s in the distribution of a form, where the differences between the different addresses of the form under consideration will be the smallest. One of the consequences of this choice is that one form can give rise to several "bundles".

Diagram 1

[1] The study included different TV programs but the detailed analyses are taken from “l’Heure de verité”. M. Brugidou: *L'election presidentielle: discours et enjeux politiques*, Paris, L'Harmattan, 1995.



For this, the judgement in probability must not be made on the overall distribution but rather reiterated at each reoccurrence of the form under consideration. It is not so much to check whether the form has a generally regular distribution, but rather to observe whether in a given part of the text, the sub-frequency of a form is remarkable given its overall frequency in the text and this as many times as the form appears[2].

1. 3. The plan of experience

1. 3. 1. Graphic forms

Built by the TROPES software, diagram 2 represents all the bundles identified in a non-directive interview using the algorithm presented previously. All the graphic forms have been retained and grammatical words)[3]. We should note that the notion of the graphic form, neither prejudges the syntactic category of the term, nor its meaning. This presents considerable disadvantages: the forms are not lemmatized (the distinction between plural and singular, the different forms marked by a verb, etc.), and furthermore homonyms are not indicated (e.g. the French word "livre" may have any one of several meanings: a book, a unit of weight, pound sterling, or the third person of the verb "livrer").

Without entering the debate over the pros and cons of lemmatizing (Salem, 1987; Tournier, 1980), it is worth reminding ourselves that by using the graphic form, we temporarily refuse to interpret the text. We are not attempting to anticipate meaning: lemmatizing would be to interfere in a process which is inherently unpredictable. To remove the ambiguity from the forms (the unit of weight *vs.* the currency) is clearly necessary, where our understanding may be compromised. Unable then to fix a clear boundary, we must therefore temporarily abandon the search for meaning. This decision is a compromise solution which enables us to advance.

The diagram reads from left to right and from top to bottom: from the first occurrence of the text to the 7639th which marks the end of the interview. The form “troisieme” (here meaning the 'third year' in secondary school) is the first bundle of the interview. It occurs four times grouped together in this first section of the text. The

---

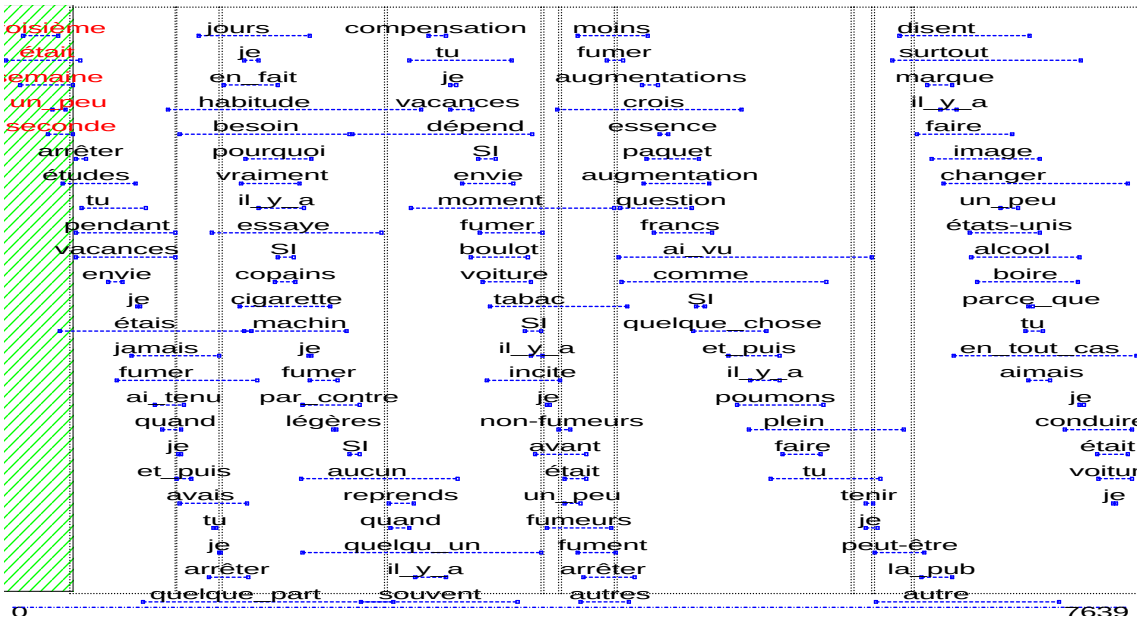
[2] It is easy to imagine the case of a form repeated often at the beginning and reappearing later - at quite regular intervals - in the rest of the text. An estimation of the overall distribution would not point out the localized concentration of the form.

[3] Given the nature of the test used in the algorithm (normal law), we will not retain the terms whose frequencies are too low (f lower than or equal to 10; for lower frequencies Poisson distribution should be used).

interviewee states that his consumption of cigarettes increased in “troisieme”, then in “seconde”, a term repeated five times at the beginning of the interview. We are thus able to follow the bundles symbolized by discontinuous lines with a dot at each end (first and last occurrence[4]) until the end of the text “je” ('I' in English). The interview finishes with the tale of a car accident provoked by smoking: "D'autant plus que je m'étais planté avec ma voiture cette fois" (*“All the more since this time I'd crashed my car").*

Diagram 2

Diagram 2: Bundles and episodes (Base: all graphic forms)



Moreover we may note that certain forms form the basis of several bundles: “fumer” - 'to smoke' -(here distinct from the third person plural form “fument”) recurs three times up to address 4181 while its use in the second half of the interview is noticeably reduced. The first person singular recurs up to ten times over the course of the interview. Thus, the personal pronoun of the first person singular, while appearing very frequently in this type of text (472 times here) does not have what can really be called regular distribution.

The diagram also illustrates that the use of the third person plural of the verb "fumer" corresponds to the only moment of the interview where the first person singular does not come in a bundle. Although this may appear unimportant, it allows us to characterize a passage in the middle of the text in which the interviewee talks about face-to-face confrontations between the smoker and non-smokers. Here the “subject” of the narrative becomes collective, proof that the absence of lemmatization can be illuminating.
In the same vein, we can see a difference between the singular and plural use of “augmentation” (increase) positioned closely within the text. This difference would appear to translate a movement of generalization of thought where there is a shift from multiple “augmentations”, each one being therefore specific, to the generic category of “l’augmentation” (the increase). In short, by induction, the subject is generalizing about his experience.

Finally, we can see that recurrence is limited (in fact, repetition leads to a sort of mechanical regularity, although the example of the personal pronoun “je” ('I') qualifies this observation). Few bundles recur: the verb “arrêter” ('to stop') appears three times during the interview (7 occurrences, then 4 then again 7).

Given that it is the central theme of the interview, it is suprising therefore to find only one bundle of "cigarette". This illustrates perhaps one of the weaknesses of the lexicometric approach. It neglects plural forms, while distinguishing between "clopes" and "cigarettes". Similarly, "tabac" (tobacco) or "legeres" (low-tar cigarettes)

[4] The wording of the bundle is located close to the bundle's address. We thus note that the bundle formed by "etais" (was) has 10 occurrences. The software identifies the address of the first occurrence -380- and the last -1673-. It enables us to edit the contexts of each occurrence and to “navigate” in the text.

can be considered, if not as synonyms then at least as "equivalents".

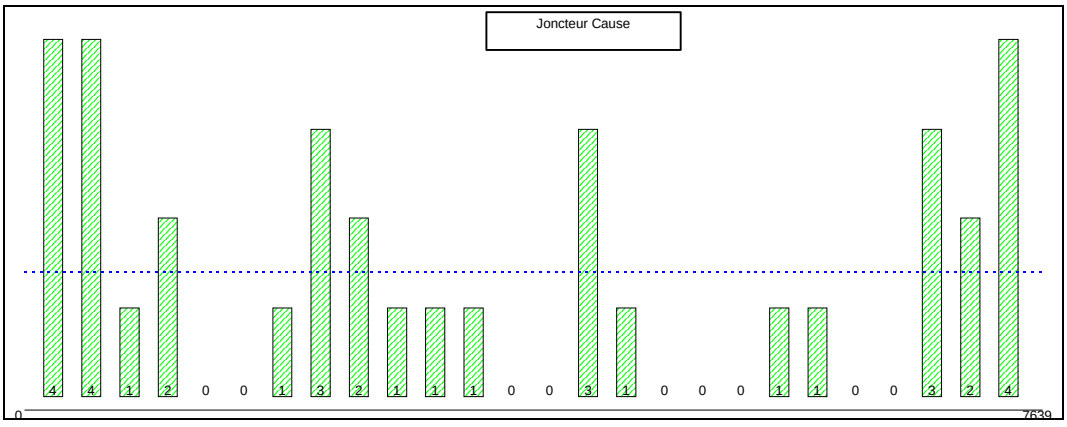
1. 3. 2. Classes of equivalent

The TROPES software enables us to go beyond the lexicometric approach by taking into account, on the one hand, certain morphosyntactic aspects, and on the other hand, by proposing a semantic analysis of substantives. It can be useful to classify grammatical words into different types of conjunctions (addition, disjunction, cause, condition, goal etc.) categories of modalization (negation, intensity, time, place etc.), or more typically into different forms of pronouns, in particular for the analysis of thematic sequences (to highlight the construction of the episodes of the narrative).
Causal conjunctions (therefore/so, because, since etc.) are a good example. They are used extensively at the beginning of the interview, when the inquiry exposes the different factors which led the narrator to smoke. We thus observe the following sequences at the very beginning of the interview:

*"Ben, aucun dans la famille avec une petite nuance tout de même **puisque** si je m'en souviens c'est que cela m'a marque." (Well, nobody in the family - although not quite because if I remember well it struck me)*
*"J'avais vu mon frangin qui avait **donc** quatre ans de plus que moi et..." (I'd seen my brother who was **then** four years older than me…)*
*"**donc** à l'epoque il devait avoir 16/17 ans," (**So** at the time he must have been 16 or 17…)*
*" je me marrais interieurement **parce que** j'avais déjà fume" (I laughed inside **because** I'd already smoked)*
*"**Donc**,ça c'etait au tout debut," (**So** that was right at the beginning…)*
*"**donc**,ensuite,en troisieme,j'ai augmente." (**So** in the third year I smoked more…)*
*"Déjà **parce que** je me suis libere un peu," (Already **because** I had some freedom)*
*"Oui, **don**c, c'etait pas vraiment du stress," (Yes, **so**, it wasn't really stress…)*
*"j'avais un peu un decalage vis-à-vis des autres **donc**," (**so** I was a bit out of step with the others…)*

TROPES will consider the forms "donc", "puisque", "car" (which can be roughly translated as, respectively, 'so', 'since' and 'because') as functionally equivalent. A long bundle of 14 occurrences of 'causal conjunctions' was thus identified before the bundle "famille"[5] located at the start of the interview.

Diagram 3: distribution of causal conjunctions per 300 word block



We can see, furthermore, that the use of negation modalizations increases significantly after the first third of the narrative when the critical remarks concerning smoking multiply *("Mais les vrais copains ne fument **pas**" - "but real friends **don't** smoke").*

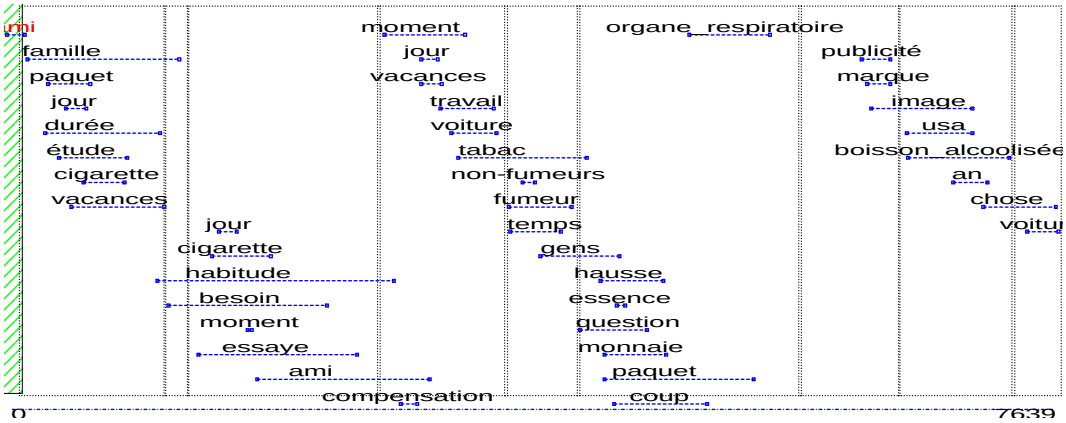
TROPES constructs classes of semantic equivalence. These classes all resemble hyponyms of a term: for example "famille", "parents", "frangin" (colloquial for 'brother') etc. are considered as equivalents of /famille/. The software proposes three levels of classification for the "references", from the widest (sphere 1) to the narrowest ("reference"). Here we will consider only the narrowest level. The construction of the "script" allows

[5] For reasons of readability, we have not included bundles formed from grammatical words on the diagrams.

us to build our own classes of semantic equivalence “by hand”, wherever we can, by pushing the initial logic of the software which then prioritizes the references within the substantives, and includes adjectives, verbs, etc.

Diagram 4 presents the diagram of bundles formed from the classes of semantic equivalence proposed by TROPES. The comparison with the preceding diagram of bundles (cf. diagram 2) constructed from all the graphic forms, highlights certain differences linked to the semantic approach used by the software.

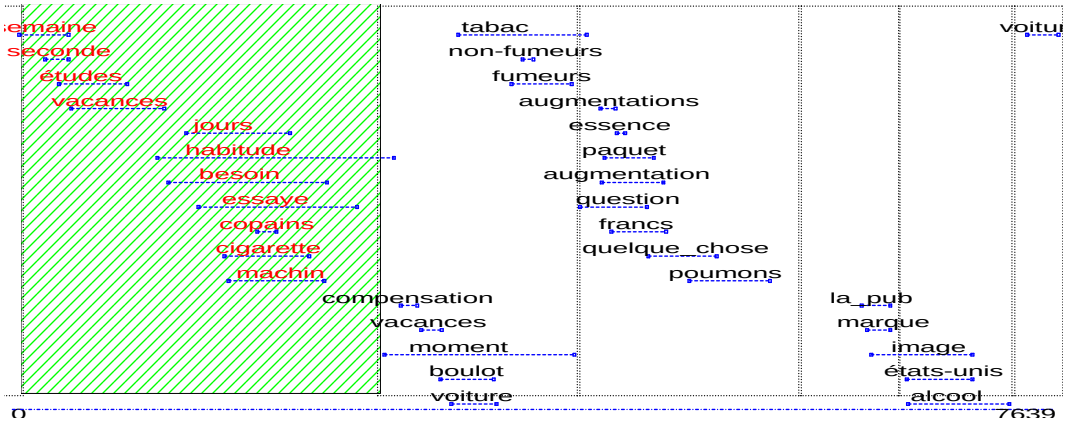
Diagram 4: Bundles and episodes (Base: Classes of equivalence -reference)



Only the classes of semantic equivalence formed from substantives appear on this diagram. Adjectives and verbs are not considered in the analysis. Grammatical words are grouped according to the principle outlined above but they are not presented in this diagram[6]. The merit of this approach (selection of substantives and semantic equivalence only) is that it renders the representation of the text clearer. At the same time, it would appear to notably reduce the quantity of information. Detailed analysis of the diagram enables us to qualify this initial observation. If we are to consider the substantives alone, the semantic approach would appear on the contrary, to be richer. The following diagram, constructed from graphic forms, illustrates this.

Diagram 5

Diagram 5 : Bundles and episodes (Base: graphic forms - substantives only)



The semantic approach gives us the bundle "ami" (friend), formed from "*copain*" (colloquial for friend), in both plural and singular form[7]. This is due more to lemmatization than semantic analysis. Its usefulness is clearly

6 The bundles are also therefore calculated according to syntactic classes of equivalent. They can be made to appear on the diagram. Nevertheless for reasons of readability, these bundles are not represented here.

7 “ T'est_entraîne par des copains et t'as envie d'essayer pour voir ce_que ça donne et_puis…” ('You're influenced by your mates and you want to try just to see what it's like and then…')
“ y a des copains ” ('there're your friends…')
“ il_y_a un copain qui m'a_propose une certaine marque et depuis ” ('a friend offered me a particular brand and since then…')

visible however since it enhances our understanding of the place of "copains" in the smoker's narrative. The same can be said for "paquet" (packet). However the presence of "famille"[8], "jour" (day) or "cigarette", included in the terms *"cigarette(s)"* and *"clope(s)"*, can be explained by semantic equivalence.

The finished canvas would appear to be more complete, giving a role to the leading players (friends, family...), evoking "cigarette". Here too our remarks need to be qualified. We have seen that most differences are due to problems of lemmatization. An interesting example is the variation between "clope"[9] and "cigarette". Moreover, the semantic approach would seem at times to be approximative. Thus the bundle "durée" ('length of time') groups together "*heures*" ('hours') of the day and "*seconde*" in the same category, mixing the unit of time and the class, "*seconde*". The "scripts" however, enable us to improve, even correct, these classifications.

Finally, we observe once more that the second sequence reconstructed by the two diagrams are remarkably similar: all the bundles are common with the exception of "machin" (colloquial for 'a thing') for the 'formal' approach and "moment" for the semantic analysis. Similar differences and confusion exist throughout the interview. We can also see that the semantic approach is better adapted to taking the less structured part of the interview into account. This is why we are now going to look at the "concurrence" of bundles, i.e., the formation of the sequence in the text via the notion of an "episode".

2. "Episodes"

2. 1. The unity of narrative meaning

For many, the sequence, or episode, represents the "missing link" in the analysis of the narrative in its longitudinal or syntagmatic dimension. It is here that non-directive interviews are an indispensable tool. The process highlights the groupings, in which bundles appear numerous and inter-linked, thus forming units of coherent text. The totality of these bundles work towards the evocation of a theme.

The graphic representation of the distribution of the bundles in the text points out the existence of a relative "vacuum". Furthermore, the presence of conjunctions (temporal in particular: before, after, then…), leads us to suppose that at these moments, the narrator is in the process of exhausting the subject, a moment of his life, and is searching for a way to "switch" to another. We lack the means to objectives this shift with any degree of certainty, unless it introduces a cut-off point where in reality there is a real attempt to create continuity. The authors of the software developed a procedure allowing us to objectivize the sequence. It consists of processing the bundles, or repetitions, which enables the text to be cut into episodes, an episode being a unit corresponding to *"the point where bundles begin and end"* (1998, p. 83).

It is primarily therefore the congruence of certain bundles which form what R. Ghiglione & al. call an "argumentative block", possessing its own referential coherence. Here in effect the bundle is no more than a visible formal indicator of all that is woven in with the referents: relationships the narrator attemps to build between the characters, events and situations that he describes. The technical contribution of the team of "cognitivists" is however more complete and more complex. In particular, they take into account the presence of "argumentative operators" (e.g. modalization of time or space) to determine, despite the prolonging or overlapping of one or several bundles, this "passage" from one episode to another.

Finally, the episode takes into account a "micro-sphere" of evocations, with its own internal coherence. This sphere can communicate with others in different ways: chronological succession, causality, metaphorical

---

" je me laissais plus_ou_moins entraîner par des copains. " ('I let myself be more or less influenced by my mates')

[8] "Aucun dans ma famille," ('Nobody in my family')
"Ben, aucun dans la famille avec une petite nuance tout_de_même puisque si je m'en souviens c'est que cela m'a_marqué." ('Well, nobody in my family - although not quite because if I remember well it struck me')
"J'avais_vu mon frangin qui avait donc quatre ans de plus_que moi et" ('I'd seen my brother who was then four years older than me and…')
" En seconde, mes parents ont_commencé à le savoir et j'ai_augmenté franchement ". ('In the fourth year my parents started to catch on and I increased alot')
" parce_qu'avec mes parents, on s'entendait franchement pas." ('because I really didn't get on with my parents')

[9] "Clopes" (slang for cigarettes) is often used in a negtative context: "on a plus de clopes" ('we've got no more "clopes"). Moreover, these terms translate distinct levels of language, where the register shifts from colloquial to formal and back again.

redundancy, etc. What is remarkable, in any case, is the mechanism of successive integrations which enables the method to be updated. Here the cognitivists' proposition coincides with that of structural analysis developed notably by C. Bremond, for whom *"elementary sequences combine to form complex sequences"* (1966, p. 61). In other words, the initial 'stammering' of the discourse, which is simply a search for and production of coherence at the level of the proposition, produces by integration a superior scale of meaning. The narrative thus appears as an unfolding plot (articulation-integration) of several episodes, each containing a part of the meaning the narrator plans to produce.

The experimental work we will consider here, consists of testing the operability of this identification procedure. We will also explore its meaning, particularly in the context of a sociological approach. A doubt remains concerning the significance of the episode, and what it may contribute to our 'understanding' of the narrative, and the intentions of the narrator. Our interrogation is due to the fact that this method can be applied to any text.
Indeed, analysis of Gustav Flaubert's classic, *Madame Bovary*, points out the existence of bundles which in turn enable us to identify the episodes. These episodes do not necessarily coincide with the organization of the novel by chapter, for example. Given that the rules of literary composition have little in common with those of narrative, what is the significance of the repetitions in this context and what is the significance of the episodes that the procedure identifies?

This examination will look therefore at the "level", to quote R. Barthes, that we consider to define the episode, based on the hypothesis that a narrative is essentially a series of interlocking sequences implying quite different levels of reference.

2. 2. The algorithm

It is via the mode of identification of the episode (or the sequence) that TROPES enables us to resolve a certain number of difficulties.

Our aim here is to dispose of a reliable method, capable of identifying these sequences and their linking while remaining 'as close as possible' to the text. We will use therefore a formal representation of the text (that given by the algorithm of the bundles) and attempt to identify co-occurrences of bundles, i.e., the parts of the text where the concentration of bundles is significant. At the same time, we hope to be able to identify thematic ruptures, those passages where the thread of the narrative unravels as other melodic lines begin.
The difficulty lies in this overlapping of sequences, and its degree of concentration. Ruptures can be strong or weak. A formal approach is therefore necessary to follow the "respiration" of the narrative. Considered one by one, each bundle seems, in its rapid *staccato*, to give the text its own rhythm. And yet their interweaving in packets creates deeper and more muted echoes. It is this, the narrative's rhythmic thread, that we must find.

Let us suppose that a new episode has begun when a large number of bundles begin and end. This treatment can be used on all four levels of classification of the reference available in Tropes (i.e. Spheres of reference 1 and 2, References used, Scenario). In all cases, certain A.P.D. (Analyse propositionel de discours, a method developed by *Ghiglione et al*) meta-categories (performative verbs, conjunctions, modalizations and personnel pronouns) are used.

When all the episodes have been identified, TROPES uses the bundle *addresses* (the average number of occurrences of the words included) to give the bundles their episodes of *belonging*. In the bundles diagram, the wording of each bundle is centered on its *address*. Although the bundles always *participate* in the construction of an episode, an episode may be empty if it contains no bundle address (the episode contains bundles which 'cross' it, but which do not 'belong' to it). Since the episodes are defined by the *edges* (beginning and end) of the bundles, and not their addresses, this must not be considered as an anomaly.

Concerning the stability of the partitioning: given that several levels of classification of the reference may be used, the number of episodes depends on the desired level of generality.

2. 3. The level of experience

The validity of the episodes proposed by TROPES must still be verified. We will test therefore the stability of the classes of bundles, notably by varying the nature of the bundle (graphic or reference form?). This validation is purely formal however; the validity of the proposed sequences must still be demonstrated from the point of view

of ‘meaning’.

The episodes create an ‘equivalent’ between the different objects of the discourse. They represent an important moment in the integration of the narrative. Our aim here is to build a transition between the level of formal description of the text represented by the bundles and the episodes and their level of description proposed by the structural analysis of the narrative.

2. 3. 1. Stability of the classes of bundles.

A simple method for ensuring the validity of the episodes is to test the internal coherence of the method. In other words, if we vary certain ‘technical’ parameters, like for example the type of bundles considered do we observe stability in the classes of bundles, i.e., the episodes? If the partitioning proposed by the algorithm is very different, the method is invalid. If it varies little, we are led to conclude that there is a certain consistence in the ‘phenomenon observed’. Given the specificity of the algorithm of the episodes, it is essentially the number of bundles taken into account in the analysis which determines the partitioning of the text. In other words, any changes liable to modify the number of bundles are important. If only the substantives or all the forms of words are considered, this parameter changes. Similarly, if we favor an approach by class of equivalence (semantic and syntactic for the grammatical words), the number of bundles changes[10].

These two parameters can of course be crossed:

	Graphic forms	Classes of semantic equivalence on substantives only
Substantives only and grammatical words by class	***A*** ***7 episodes*** ***diagram 5***	***B*** ***10 episodes*** ***diagram 4***
All syntactic categories (grammatical words are not in classes)	***C*** ***12 episodes*** ***diagram 2***	***D*** ***12 episodes*** ***diagram 6***

The experiment has been simplified here:

- The lines show either the substantives and the grammatical words in classes (types of conjunctions, moralization etc.), or all the syntactic categories (substantives, adjectives, verbs... and the grammatical words which are not grouped in classes),

- In the columns, the alternative between the graphic forms and the classes of semantic equivalence is tested.

By neutralizing the ‘grammatical words in classes’ variable we note an important effect linked to the classes of semantic equivalence. The algorithm identifies ten episodes (B) and not seven (A). However, if we consider all the syntactic categories, the difference linked to the groups of substantives in classes of equivalence is zero, in quantitative terms (twelve episodes in C and D). We know therefore that ‘qualitatively’, the semantic approach produces certain effects. By neutralizing the graphic forms variable we note that if we consider all types of word the number of episodes increases significantly (from seven -A-, to twelve -C-). Conversely, if we consider all the syntactic categories the increase is smaller (ten -B- to twelve -D-).

This effect must be viewed in a different light: by considering both the number of episodes and the definition of these episodes. Returning to the diagram confirms the stability of the proposed groupings.

[10] To be complete, we should add that all manipulations of frequency thresholds have an influence on the number of bundles. For example, if we only consider bundles longer than 3, 4 or 5 etc., but also if we only consider bundles formed from words whose frequency is above a certain threshold (here above or equal to 10).

In effect, the partitioning into episodes proposed in option A (cf. diagram 5) is not questioned by the other options but rather improved. The new classes of bundle exist in effect in all cases except one[11] inside those classes already identified. In other words, the B options (taking into account of semantic equivalence within the substantives), C and D (taking into account of all words) enables us to identify all sequences inside a larger sequence that we can therefore consider as 'stable'.
The change of viewpoint enables us to ensure the stability of the episodes which are not here artifacts due to the method and to enrich the interpretation of the episodes by identifying sequences inside the episodes identified. We thus highlight the 'Russian doll' nature of the narrative.

Moreover these comparisons are interesting because they outline the first stages of an ***interpretative method*** based on the following two stages:

Verification of the validity of classes of equivalence on the substantives ==> Identification of sub-sequences in the episode, distinction between the rheme and the theme through the study of other types of word

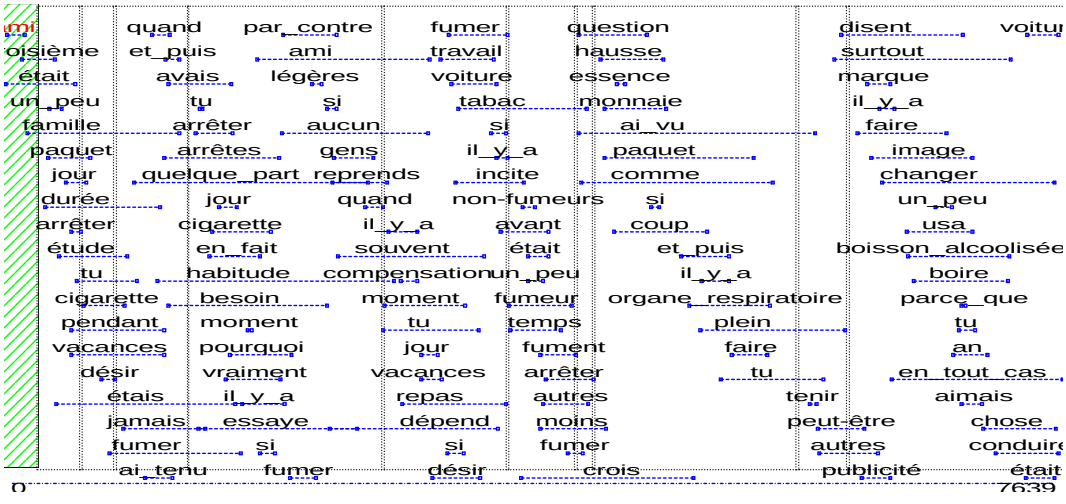
2. 3. 2. The theme and the rheme

The comparison of the different options for analyzing the text draws our attention to another phenomenon of fundamental importance. We have seen that the taking into account all types of word (*as opposed to* substantives only) does not dramatically modify the partitioning of the narrative into an episode. This description is simply more precise. We will observe nevertheless, that the distribution of the substantives by bundle and the distribution of the 'other words' (adjectives, verbs, grammatical words...) is not identical.

Let us consider diagram 6 (or 2) and compare it to diagram 4 (or 5). We can see that there is a significant concentration of the 'other words' between the groups of bundles formed by the substantives.

Diagram 6

Diagram 6: Bundles and episodes (Base: classes of equivalence and graphic forms on non-substantives)



In other words, between each large grand episode identified, we can see a series of verbs and grammatical words. Thus between episodes 2 and 4 of diagram 4, we observe a series of 10 terms (*etais (was), jamais (never), fumer (to smoke), ai tenu ((I) held), quand (when), et puis (and then), avais (had), arrêter (to stop), arrêtes (stop), quelque part (somehow/somewhere*) positioned between the last substantive of episode 2 ("vacances") and the first of sequence 4 ("jour"). The software therefore identifies a third episode formed solely from grammatical words.

[11] Note that taking into account all words (vs. substantives only) erases the last sequence based around "voiture" (car) (passage from A and B to C and D).

Similarly, between episodes 4 and 5, we identify a series of six 'other words' (*legeres (lights), aucun (none), reprends (start again), quand (when), il y a (there is/are), souvent (often)*) positioned between "ami" and "compensation". The phenomenon is not as clear between episodes 5 and 6 (only 2 terms) unlike between 6 and 7 where we count six terms: *fument (they smoke), arrêter, autre (other), moins (less), fumer, crois (believe/think),* etc.; and between episodes 7 and 8 (five terms). The last two transitions between sequences are less clearly marked by such concentrations of 'non-substantives'.

It is now possible to reconstitute the construction of episodes by distinguishing between the reference (theme or topic) and the commentary (rheme). From the bundles diagrams, we will therefore make a first schematization of the narrative. In the following table, we have considered only the largest episodes (cf. diagrams 4 and 5). In column one, we find the substantives, in column two, the other types of word.
The substantives are grouped in a very large class of equivalents. In sequence A for example: "ami" (friend) and "famille" indicate people; "paquet" (packet) and "cigarette" refer to the same object; "etudes" (studies), "jour" (day), "vacances" (holidays) refer to periods of time. The provisional classifications identify the narrative functions at work in the text. Secondly, we must characterize the type of relation that exists between these different objects by identifying the narrative functions. Here we assume functional equivalence between these objects. Only the level of formalization proposed by the structural analysis enables us to confirm this.

Furthermore, between sequences, we have identified transitions between episodes, passages where the narrative flows primarily through the 'commentary'.

Table 1

THEME (see diagram 4)	**RHEME (see diagram 2)**
Sequence A: ami paquet etudes famille cigarette jour vacances	*un peu, pendant* *arrêter*
Transition A/B	*jamais, quand, et puis, quelque part* *etais/avais, arrêter, arrête, ai tenu, fumer*
Sequence B: ami cigarette jour habitude moment besoin	*en fait, pourquoi, vraiment, par contre* *il y a, fumer*
Transition B/C	*legeres, aucun, souvent* *reprends*
Sequence C: tabac moment travail compensation jour voiture vacance	*depends, fumer*
Transition B/D	*il y a, incite*
Sequence D: non-fumeurs temps fumeurs	*avant, un peu,* *etait*
Transition D/E	*autres, moins,* *fument, fumer, arrêter, crois*
Sequence E: poumons paquet hausse essence question monnaie	*comme, et puis, il y a, ai vu*
Transition F/E	*plein (?), peut-être, autres,* *faire, tenir*
Sequence F: publicite usa alcool marque image	*un peu,surtout, parce que,* *il y a, disent, faire, boire*
Transition F/G	*en tout cas,* *aimais*
Sequence G: voiture	*etait, conduire*

This table indicates the levels of transformation between the different episodes of the smoker's narrative.

We note for example strong continuity between sequences A and B. Certain bundles disappear (famille, paquet, etude, vacances) although the structure of the theme does not appear to be significantly modified. The transition between the two sequences (passage A/B) constitutes a variation on the verbs “arrêter” , “fumer” and “tenir” (respectively to stop, smoke and hold).
Here, the narrative ‘changes foot’, no longer evoking the origins of the smoker's history (imitating friends, facing up to his family), but rather the realization that a habit is setting in. The beginning of two new bundles, “habitude” (habit) and “besoin” (need), indicates a new melodic line which overlaps the first, sometimes modifying it profoundly.
Returning to the text makes the different mechanisms of the unfolding story clear, identifying the actors and the forces at work behind the scenes. The concepts of narrative structural analysis must now take over: the representation of the proposed text must constitute the first substrata of this analysis by indicating the principal sequences, and identifying the linking and the ruptures.

Conclusion

The experiment confirms the relative stability of the episode objectivized, in the order of the narrative, by the bundle method. The program identifies this minimal unit around which the narrative is built, and also brings out the logic of integration of its different episodes. This work is intended as one of the possible methodological translations of the propositions, made notably by R. Barthes at the end of the 1960s, which aimed to found a structural analysis of narrative. It is based on the two fundamental functions implied by all narratives: the integration of secondary episodes in the broader meta-episodes, and the articulation of these meta-episodes.

The utility of identifying the episode, and its constitutive logic, has been confirmed by members of the cognitive sciences. The contribution of the ‘words’ research group is particularly influential here, in that it demonstrates the fundamental mechanism of the search for referential coherence, and the limits of memorization. But these arguments, though explanative, remain largely incomprehensible for the interpretive sociologist who must focus on the ‘target meaning’ of the narrator, in the sense given by M. Weber to this expression.

Bibliography

Ouvrages:

Babayou P. : *Traitement des questions ouvertes : comparaison d’une post-codification et de méthode lexicométrique et d’analyse du discours*, Paris, Cahiers de recherche du CREDOC, n° 101, septembre 1997.
Blanchet A., Ghiglione R. & al. : *Les techniques d’enquête en sciences sociales*, Paris, Bordas, 1987.
Brugidou M. : *L’élection présidentielle : discours et enjeux politiques,* Paris, L’Harmattan, 1995.
Demaziere D. et Dubar C. : *Analyser les entretiens biographiques, l’exemple des récits d’insertion*, Paris, Nathan, 1997
Ghiglione R. & al. : *L’analyse cognitivo-discursive*, Grenoble, Presses Universitaires de Grenoble, 1995 ; *L’analyse automatique des contenus*, Paris, Dunod, 1998.
Labbe D. : *Le vocabulaire de F. Mitterrand*, Paris, Presses de la Fondation Nationale des Sciences Politiques, 1990.
Lafon P. : *Dépouillement et statistiques en lexicométrie*, Paris-Geneve, Slatkine-Champion, 1983.
Muller C. : *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette Universite, 1973 ; *Principes et methodes des statistiques lexicales*, Paris, Hachette, 1977.
Salem A. : *Pratique des segments répétés*, Paris, Meridiens Klincksieck, 1987.
Schutz A. : “ Phénoménologie et sciences sociales ”, in *Le chercheur et le quotidien*, Paris, Meridiens-Klincksieck, 1987.

Articles

Brugidou M. & Le Queau P. : “ L’analyse des entretiens non directifs par la methode des rafales ”, *JADT*, Giornate internazionali di dati testuali, Rome, CISU, CNR, 1995.
Hubert P. & Labbe D. : “ La repartition des mots dans le vocabulaire presidentiel (1981-1988), *Mots*, n° 22, mars 1990.
Jenny J. : “ Methodes et pratiques formalisees d’analyse de contenu et de discours dans la recherche sociologique française contemporaine ”, *BMS*, n° 54, 1997.
Labbe D. : “ Des reformes à la cohabitation, les quatre périodes du premier septennat Mitterrand ”, *Mots*, n° 22, mars 1990.
Lafon P. : “ Statistique des localisations des formes d’un text ”, *Mots*, n°2, mars 1981.
Tournier M. : “ D’où viennent les fréquences de vocabulaire? Le lexicométrie et ses modèles ”, *Mots*, n° 1, 1980.