TRIAL AND CRITIQUE OF ALCESTE AS A TOOL FOR ANALYZING SEMI-STRUCTURED INTERVIEWS IN SOCIOLOGY

Monique Dalud-Vincent

Maison des sciences de l’homme | « Langage et société »

2011/1 No 135 | pages 9 - 28

ISSN 0181-4095
ISBN 9782735113200

This document is the English version of:

DOI 10.3917/ls.135.0009

Translated from the French by JPD Systems

Available online at:


How to cite this article:

DOI 10.3917/ls.135.0009

Electronic distribution by Cairn on behalf of Maison des sciences de l’homme.
© Maison des sciences de l’homme. All rights reserved for all countries.
Reproducing this article (including by photocopying) is only authorized in accordance with the general terms and conditions of use for the website, or with the general terms and conditions of the license held by your institution, where applicable. Any other reproduction, in full or in part, or storage in a database, in any form and by any means whatsoever is strictly prohibited without the prior written consent of the publisher, except where permitted under French law.
Trial and Critique of Alceste as a Tool for Analyzing Semi-Structured Interviews in Sociology

Monique Dalud-Vincent
Université de Lyon
MEPS- Centre Max Weber
Monique.vincent-dalud@univ-lyon2.fr

Abstract
As the number of textual analysis software programs increase, sociology is faced with the problem of gauging their suitability. On one hand, they save time insofar as they can process a large volume of texts but, on the other hand, the way in which the data is analyzed and the default hypotheses implemented in these tools require consideration. In this paper, we focus on the Alceste program and, based on a particular set of problems, attempt to deconstruct the way in which it processes text in general and semi-structured interviews, in particular. Reinert constructs a “summary” (among many possible alternatives) of the corpus considered as a whole by taking into account mainly hypotheses on words that concentrate the meaning of a text according to their grammatical category, but also hypotheses on text partitioning into Context Units, while seeking to remain within the “proper” conditions for applying certain statistical tools. This summary and analysis, based on regularities, do not, at any point, incorporate the sociologist’s hypotheses or theoretical viewpoint.

1. My sincere thanks to Sylvia Faure and Marie-Carmen Garcia, members of MEPS-Centre Max Weber, for allowing me to analyze their corpus for this paper.
Alceste\(^2\) is a statistical analysis tool for textual data. While its creator, sociolinguist Max Reinert, mainly used Alceste for analyzing poetry and literary texts (Reinert 1990a, 1990b, 2000b, 2001a), the documentation states: “Alceste is a software tool developed to assist in the analysis of a textual corpus: interviews, responses to open questions, literary texts—in fact, any document written using the Latin alphabet, the ten numerical digits, and the usual punctuation signs, provided there is a certain amount of homogeneity and minimum volume” (Reinert 2000a). It is possible, therefore, to work not only with interviews, but also with other documents such as newspaper articles and legal documents.

Based on a sociological aim and a set of interviews on the subject of hip-hop dance (Faure 2004; Faure and Garcia 2002; 2003; 2005), this paper aims to address, in part at least, the following questions: Can this tool contribute something? If so, what is its contribution? The purpose here is not, however, to come down in favor of either using Alceste or carrying out the work by hand (i.e. without any special tool), but rather to reflect upon what we are doing/not doing when we analyze interviews, and to take a step back and observe a machine whose functions are based on precise actions, divisions, and so on, and which cannot truly interpret the meaning of the language it is analyzing.

First, we will describe how Alceste works. This will allow us, at the same time, to discuss what this software is not able to do inasmuch as its objective is limited. Second, we will present results from an Alceste analysis of the hip-hop corpus. We will then show the various choices sociologists will need to consider in light of the effect these choices will have on the software’s results.

1. **Alceste: Theory, Functions, and a priori Limitations**

1.1. **The Summary Table**

Max Reinert was a student of Jean-Paul Benzécri, considered to be the founder of so-called French style data analysis. It is not surprising, then, that Reinert turned to the statistical analysis of his texts with a whole panoply of tools, such as the chi-square test of association, correspondence analysis (CA), agglomerative and divisive hierarchical cluster analysis (AHC and DHC), and particularly the pre-formatting of the corpus in a two-dimensional table for analysis.

Generally speaking, Alceste’s methodology consists of studying the laws of vocabulary distribution in a corpus. It is concerned “not with

\(^2\) ALCESTE stands for Analyse des Léxèmes Cooccurrents dans les Énoncés Simples d’un Texte (Analysis of co-occurring lexemes within simple utterances in a text).
finding the meaning of a text, but with determining how the elements
that make up the text are organized” (Reinert 1990b).

What constitutes a corpus as far as Alceste is concerned? It is a single
document containing all of the interviews: “A corpus will be defined as
any group of texts assembled by someone for the purpose of analysis, for
example, a group of articles on a given theme, a group of responses to
the same question, etc.” (Reinert 2001b).

The author stresses the fact that, in order to obtain a significant result,
two conditions must be met (Reinert 2000a):
– the corpus is presented in its entirety and as having a certain coherence
(a thematic coherence or a coherence in terms of production conditions)
– the corpus must be big enough to be able to carry out a viable statistical
analysis. The idea of Alceste is to produce an overall picture of a very large
document very quickly.

The goal is, therefore, not to deal with each interview separately but
rather to consider the group of interviews as a coherent whole.

More precisely, Alceste disregards certain words (function words) and
conjugation endings (its focus is on the roots of words) and preserves
content words (verbs, adjectives, nouns, and adverbs). Function words
(markers of time, space, etc.) may, of course, serve an illustrative purpose
even if they do not play an active role in the treatment of the text. The

3. The file must be saved in ASCII format. It is partitioned into Initial Context Units
(ICUs) representing the different interviews. Each ICU is introduced by a line con-
taining variables that serve as examples.

4. The user need not be involved in this lemmatization and classification of function
and content words. Alceste uses “integrated” dictionaries, which are able to convert a
French plural noun into its masculine singular form, verbs into their infinitive forms,
and so on. It is interesting to delve into this level of detail in Alceste’s operations.
Under the heading of the French verb taire, for example, the word tait is listed. This
third person singular form of the present tense of the verb taire is a homonym for
“tait,” as in “c’était,” a familiar, shortened form of the phrase “c’était,” or “it was.” This
illustrates the kind of difficulty posed by spoken language as it is typically transcribed
by sociologists.

This partitioning is doubly functional insofar as it is grammatically functional
(for the marking up of content words, which enables grammatical catego-
rization) and mathematically functional in the sense that the content
words, \(^5\) once reduced, will constitute the columns of the table necessary for a Benzécri-style analysis.

After partitioning, it follows, therefore, that the rows of the table will also be filled in. These will be made up of the partitions of the corpus as utterances (i.e. parts of the discourse or text). Following Benzécri’s logic, this will make it possible to construct a presence/absence table, filled out according to whether or not the reduced content word (i.e. root form) is present in the text. The corollary question is therefore: what constitutes an utterance? Is it a phrase, a sentence, a paragraph, an interview, or some other unit?

It is worth noting in passing that sociologists sometimes produce presence/absence tables when crosstabulating interview subjects with the various themes considered important for analysis.\(^6\) This type of analysis makes it possible to classify the interviews by grouping them according to themes, for example.

Alceste (Reinert 1993; 1995) goes further in terms of partitioning: first, in relation to the columns, as we saw above (identifying content words rather than themes)\(^7\) and, second, in relation to the rows, by defining an utterance as a small part of an interview. It seems difficult, however, to incontrovertibly define what this “small part” should be.\(^8\) Several interlocking partitions are created in Alceste. The standard version constructs Elementary Context Units (ECUs, not to be confused with the Initial Context Unit, or ICU, which corresponds to a whole interview). The ECU is the basic statistical unit since Alceste classifies ECUs according to the distribution of vocabulary. The user is not required to intervene at this stage: an ECU (generally consisting of a few lines of text) is defined as a function, on the one hand, of its length in number of occurrences

\(^5\) The root (lemmatized) forms of content words that prove to be hapax legomena (i.e. occurring only once in the text) are removed. Thus, not all words will be analyzed. Regularity, not singularity, takes precedence.

\(^6\) With Alceste, the user has no real means of examining the corpus based on their own hypotheses. The analysis will be to the same configuration, irrespective of any hypotheses, and the general procedure will not be changed.

\(^7\) It would be possible to propose some themes by carrying out preliminary work to link each theme up with the words that might appear in the corpus relating to that theme. This is not within Alceste’s scope, where the potential to interpret data has been eliminated as far as possible (though, in reality, the partitioning of content words already constitutes a certain reading of the corpus and, therefore, a particular interpretation).

\(^8\) It would be fairly simple, however, to create an “interview x themes” table once a list of themes and associated words had been drawn up, which could save the analyst some time.
(of words) and, on the other, of its punctuation. This first partitioning is questionable inasmuch as the results might be different if the partitioning is carried out differently. Furthermore, an ECU will not necessarily correspond to a meaningful unit; it could just as easily consist of a fragment of a sentence as it could an entire group of several sentences.9

Nevertheless, an initial table is constructed by crosstabulating all of the ECUs in the corpus with the content words in root form (provided that the reduced form appears in at least four ECUs, which has the effect of further eliminating outliers). In each cell of the table, the digit “0” appears if the form does not occur in the ECU and a “1,” if it does.10 At this stage of the analysis, no distinction is made among ECUs from different texts or interviews, the purpose here being to work with the corpus as a whole.11

Reinert then constructs less fine-grained partitioning. Context Units (CUs) are formed by joining together consecutive ECUs until the number of root forms analyzed within the UC exceed a fixed value, k.12 Once again, no intervention is required from the user at this stage. Reinert designated two possible values for k since there was nothing to indicate that the results would not differ as a result of this parameter13 (thus, in reality, two tables are constructed, one for each of the retained values of k).

It is clear that the construction of this table (for fixed value, k), independently of any analysis that might be applied, leads to very restrictive limitations, sociologically speaking. Reinert constructs a “summary”14 (among many possible alternatives) of the corpus considered as a whole by taking into account mainly hypotheses on words that concentrate

---

9. There is nothing to say, moreover, that the ECU might not be made up of a sociologist’s question followed by an interviewee’s response (or fragment of a response).
10. The number of times the root form appears within the UCE is not taken into account.
11. The table is nevertheless limited to a maximum of 1,400 root forms (of content words) and 10,000 ECUs.
12. The number of rows in the table is decreased, therefore, and the incidences of “1” are concentrated automatically. Nevertheless, the density of 1s remains fairly unimportant. This density represents 1% in our present corpus. Indeed, certain verbs, nouns, and adjectives, which are very closely linked to the topic, appear quite frequently but, at the same time, a large number of infrequently-appearing content words (the singletons have their say!) may appear. Reinert indicated that there can be more than 90% of 0s in the cells in the table. This percentage may challenge the validity of an analysis of the regularity of the occurring forms.
13. The two automatically-generated values are relatively close to each other (in our corpus, they equaled 10 and 12).
14. Or rather, a reduction, for there is no indication that what will be preserved is in fact “essential.”
the meaning of a text according to their grammatical category, but also hypotheses on text partitioning into Context Units (considered units of meaning by Reinart), while seeking to remain within the “proper” conditions for applying certain statistical tools.

The “who is speaking?” question is removed from the analysis, for example. Already, we can see that the order of the rows (CUs) and columns (root forms of content words) does not change the “summary” because the same information is always addressed. It is also clear (and we will see this in the analyses produced by Alceste) that the internal logic of a discourse is not studied at all (even if the table displays the CUs in the “correct” order). Furthermore, each CU carries equal weighting in Alceste, meaning that each is accorded the same importance. Therefore, while the sociologist might consider part of an interview as “off topic” or secondary, Alceste will nonetheless treat it in the same manner as the rest of the text. One of our interviews, for example, took place on the patio of a café and, during the interview, an intoxicated person interrupted both physically (by bothering the participants) and verbally (the interviewer and the interviewee both spoke with this person). Although present in the transcription, this passage would not be analyzed by the sociologist, but Alceste does not make a distinction between it and the rest of the interview. This is quite an extreme example but there are other, more subtle examples such as when an interviewee sometimes speaks at length on a subject that is not the focus of analysis. With Alceste, no hierarchization within a single person’s speech is possible.

This is the case more generally in relation to our present purposes insofar as we are dealing with semi-structured interviews where the interviewer’s questions are intermingled with the interviewee’s responses (in the same interview and, thus, in the same ICU). Alceste has no means of sorting out the text and will treat it, again, as though it is all valid. The user can, of course, take action by removing the sociologist’s questions and retaining only the responses, but this choice would be contrary to the standard interpretation of an interview, which takes into account the interactive context (Beaud and Weber 2003). Moreover, how can the meaning of a response even be understood without knowing what the

15. Of course, the sociologist could sift through and remove these passages from the corpus.
16. This problem does not exist in relation to literary texts where the disparity is measured between the (theoretical and methodological) questions arising according to the material being analysed and the different objectives that are set.
17. Some ECUs will be spoken entirely by the interviewee, some by the interviewer, and some will consist of a combination of words spoken by both.
question was\textsuperscript{18}. The sociologist, therefore, has to make a choice if they wish to use Alceste as to whether they retain the interviewer’s words and commit to analyzing the text as an integrated whole.

It is also worth noting that the original forms of the words (as they appear in the interview) are lemmatized and there is, therefore, no analysis possible of the different tenses of a verb, for example, or the varying subjects of that verb,\textsuperscript{19} which are nonetheless informative sociological “readings” for the analyst. In addition, the many variants, such as affirmatives, negatives, and interrogative statements, which can co-occur with the same verbs, adjectives, nouns, and adverbs are lost in the table and, thus, from the analysis. In other words, Alceste produces the same summary for both a negative and an affirmative utterance (UC), or for a statement and a question.

Another limitation of Alceste concerns all the information lost as a result of the interviewee’s manner of speech, intonation, hesitations, self-presentation, and so on. Everything that the analyst may hear (either at the time of recording or when listening back to it afterwards, such as a silence, the tone, etc.) or see (demeanor, nervousness, etc.) is considered unimportant (or, at least, less important than the information retained for analysis) using such a summary table. Its purpose is to separate the manner of speech from its context, even though latter may give the former its meaning. What does the analyst make of a sentence such as the following from one of our interviews, especially when they listen to the recorded version: “At that time, I was THE\textsuperscript{20} person. . .”? And what do they make of a sentence\textsuperscript{21} spoken with humor by a young man from a working-class neighborhood (dressed in such a way as to leave no doubt as to his social origins), such as “My father? He’s a senior executive!”? Of course, this is to be understood ironically! Alceste would retain “senior executive” and not “working class.”

Finally, we must take into consideration everything that the sociologist can “read” (and therefore understand and analyze), but which Alceste’s various dictionaries are unable to classify without any sort of preliminary correction.\textsuperscript{22} There is, in fact, a substantial difference between spoken language and written language; the language used in an interview is much

\textsuperscript{18} It is rather strange to read an interview after the sociologist’s speech has been removed.

\textsuperscript{19} For example, the sentence “he would like to fly airplanes” would be summarized by two verbs (“like,” “fly”) and one noun (“airplane”). This summary would be exactly the same for “I like to fly airplanes.”

\textsuperscript{20} The speaker “stresses” this word, and we have represented that stress visually here.

\textsuperscript{21} This example was created for illustration purposes.

\textsuperscript{22} Can these truly be called corrections?
closer to spoken language, especially in situations where the sociologist is interviewing people from working-class areas, who are not equipped with an education. This fact is not without its consequences when using Alceste. In our interviews, which were conducted with dancers from working-class neighborhoods as well as with choreographers, we observed dramatic differences in the type of language used by these different subjects. These differences, whether they were in sentence construction, grammar, or vocabulary, are all worth studying. A simple word could flag up a distinctive feature, for example, a subject may use the French slang term meuf instead of the standard term femme—both words may be translated as “woman”—but the slang term would not be recognized by Alceste without some preliminary correction. This problem also occurs with words that the sociologist transcribes as they were pronounced in order to remain faithful to the original, for example, the French word for “little,” petit, is often transcribed in its shortened, oral form, p’tit. In order for any word in the corpus to be recognized and analyzed, the software must be able to link it to a root word from a dictionary. Alceste also requires special formatting using the underscore symbol for composite expressions or terms, such as “Maison_de_la_Danse,” “Saint_Étienne,” and “hip_hop.” The sociologist will have to decide whether to apply some or all of these types of corrections to the corpus so that the program is able to classify these words. The alternative is to choose not to use this tool (or any other) when analyzing language containing a lot of slang. This is a question that concerns not only sociological practice, but also theory.

1.2. Analyses Produced by Alceste

Alceste’s entire analysis is based on the summary table, classifying CUs using hierarchical descending classification (HDC). The CUs are grouped together based on the guiding hypothesis that two CUs belong in the

23. This is practically a moot point when dealing with a literary work or poetry written by a single author.

24. When Alceste comes across this type of writing, it replaces the apostrophe with a space, thus reading two separate words (in this case, “p” and “tit”). The user has two options: either they “correct” the word (by replacing p’tit with petit), or they link the two parts with an underscore (to make “p_tit”).

25. We must bear in mind that the fact that certain words are both nouns and adjectives (for example, the word “innocent”) does not aid the categorization process.

26. In this paper, we will not provide a detailed discussion of the methodology, as this can be found in Reinert’s publications listed in the references and Image’s publication (Image, 2007).

27. Since two k values are determined, there are actually two tables and two parallel classifications.
same class if they contain the same content words in root form. They belong in different classes, however, if they rarely or never contain the same root forms. To separate the CUs into two classes, the program starts with the discrepancy (from the chi-square value) between the two groups of CUs which constitutes one partition of all of the CUs. It identifies the two groups that are most different from each other and these make up the first division in the corpus. Depending on the quantity of classes desired (which is determined by the software, but may be adjusted), Alceste will create a second split, beginning with the class containing the most CUs. Thus, one of the initial groups is separated into two groups, and so on.

It is worth noting that, in general, a single class of CUs will contain units drawn from several interviews. Here again, no distinction is made among the various interviews in the analysis. Using Alceste to separate the corpus into classes does not mean dividing the interview subjects into classes. Rather, the result is the creation of “packets” of speech fragments which are grouped according to the root forms of the most frequently-occurring content words. Thus, a single interview is inevitably scattered among several classes, and each class may contain fragments of sentences spoken by several people.

Ultimately, two results are obtained (one dendrogram for each of the two CU “lengths” at a fixed value of k). Before interpreting the data, the two results must be compared. Alceste makes the comparison by crosstabulating the classes obtained in the two dendrograms, and identifying the “stable portion,” or the groups of ECUs of the same class, as determined by the two HDCs. In this step, it is the ECUs that are grouped, and not the CUs, which will have varied from one table to another. For each pair of classes, the program evaluates this stable portion by calculating a directional chi-square test of association between two classes: class i for the first classification, and class j for the second. This takes the form of a table containing two lines: for the first HDC, the i class is measured against the other classes combined; for the second HDC, there are two columns with the j class measured against the other classes combined.

28. The results of an HDC are represented by a dendrogram showing the different classes and the ways in which they overlap.
29. The user has still not been required to intervene in the program’s functioning at this point (unless they have chosen to adjust the program’s settings, but this choice does not alter any of the steps taken).
30. This chi-square test accepts the same formula as the one that is normally used, however it is directional, meaning that it has the same sign as the deviation from independence in the square of the table (i, j). It indicates whether the measured dependence represents an attraction or a repulsion between classes i and j.
each cell, there is the number of classified ECUs. These calculations make it possible to identify the significant, positive values of chi-square, and to preserve the stable groups of ECUs, so that they may be subsequently described and interpreted.

Consequently, ECUs that do not fall into the stable portion will be eliminated from the analysis, the idea being to only retain the ECU groups that show significant concordance between the two classifications. Reinert has stated, moreover, that, as a general rule, the stable portion exceeds 50% of the ECUs in the corpus (Reinert 2001b).

It is clear that this step imposes yet more limitations. First and foremost, there is the limitation imposed by having to choose the $k$ values and the length of a CU. Choosing alternative values (in particular, values that are very different from each other) may produce different results for the HDC. In response to this concern, Reinert used experiments to demonstrate that this choice did not fundamentally change the results (Reinert 1993), but it is possible that his conclusion may not be valid for other types of corpuses. Further, the observation that the stable portion contains at least half of the ECUs in the corpus also means that perhaps half of the ECUs (equaling half of the entire corpus) will be abandoned in the analysis as a result. This is definitely not without significance. Knowing that the “abandoned” portion is the part that did not appear in either analysis of the corpus, and which thus might appear to be “unstable” or outlying, reinforces the emphasis on regularity.

1.3. Aid to Interpretation

There has been no real need for any intervention from the user up to this point in the analysis. Had any intervention taken place, it would have been to adjust the settings (which does not alter any of steps taken by Alceste), to “correct” certain words in the corpus so that they may be recognized by the software, or even to simply format the corpus to

31. If the chi-square value is positive, the two classes ($i$ for the HDC1, $j$ for the CDH2) have an “attraction.” The higher the chi-square value is, the stronger the attraction is. The stable portion is determined by retaining the groups of ECUs in the intersections of those classes that obtained a strong, positive chi-square value (a maximum for the row and column).

32. The question of choosing the HDC as a classification tool will not be addressed here.

33. No use is made of the differences found between the two classes. Alceste only gives a list of ECUs that are “non-classified.” The proportion of non-classified ECUs varies considerably depending on whether “corrections” were made to the corpus. This leads to the question of whether it is best to make “corrections” knowing there is a risk of “enlarging” the portion of the corpus that will be analyzed.
separate the individual interviews into ICUs, and to insert a few variables into the text for illustration purposes. There is thus a swift progression from preparing the corpus to interpreting the results. At this stage, the user will see a display of columns containing the root forms (as well as the “stable” classes of ECUs). The reading and interpretation of the classes can now begin, but what should it be based on?

Reinert relies on the principle that each class corresponds to a “semantic world” or a “lexical world.” According to Reinert, it is by seeking out the internal coherence of the vocabulary of each class as well as the inter-class differences of the vocabularies that it is possible to label each class of ECUs and to observe the broad trends in the corpus.

Reinert provides complementary analyses of the data in Alceste as an aid to interpretation.

The software starts by producing statistical data on the strength of a given content word’s association (in root form) to each of the classes. This output is in the form of two tables: one showing “significant presences” and another showing “significant absences” for each class. Again, the criterion used is a directional chi-square value of association, calculated in the same way as before. This time, it is a matter of crosstabulating the rows (one containing one of the stable classes and the other containing all the other stable classes combined) with the columns (one containing a fixed content word and the other containing the ECUs that do not include that content word). Each cell displays the number of ECUs concerned. The number of ECUs in the class containing the form is also given as well as the percentage of ECUs that this represents with respect to the total ECUs containing this form across all classes.

These calculations are carried out according to the same principle for the different variables and different function words. Although these elements did not influence the analysis, Alceste shows whether their presence is significant. The program carries out the same grammatical categorizing as it did in the initial lemmatizing phase, indicating whether the concentration of content, and even function, words is significant in a given class.

34. The risk here is to give the impression that the analysis of texts and interviews is a rapid and simple process thanks to tools such as this one, which can even analyze a corpus that has never been read (in the poorest sense of the term). This is very far removed from the recommendations espoused in sociology manuals (Beaud and Weber 2003).

35. The chi-square sign is determined by the sign of the deviation from independence in the cell in which the isolated class is crosstabulated with the ECU containing the fixed content word.
A factorial correspondence analysis (FCA) rounds out the initial results, giving a graphical representation of the classes and any links they may have to each other. Alceste uses the table that crosstabulates the root forms retained in the analysis with the stable classes to do this. Each cell displays the number of ECUs in any given class containing the root form.

Alceste also gives the original forms of the words (as they appeared in the interview) for the root forms. This creates a bridge between the root forms retained in the various tables (as significant presences and absences) and the interviews, using concordancers that link to the ECU containing the form. It also shows the most representative (or significant) ECUs for each class (by calculating, according to the same principle as before, a chi-square of association called a “coefficient of association of an ECU to a class”). Finally, this makes it possible to visualize the corpus in which the ECUs exhibit the characteristics of their class.

This will be the basis upon which the user may formulate an interpretation. The interpretation cannot be formulated in absolute terms, but rather with expressions such as “more frequently” or “less frequently.” A form with a strong positive chi-square value for the stable class i should be viewed, for example, as “more frequently” present in the ECUs for the (stable) class i than for all of the (stable) classes as a whole. This does not mean the form is absent from the ECUs of other stable classes, nor does it mean that the form is absent from the ECUs that were removed. It also does not mean that this form indicates a significant presence in another class. Furthermore, the fact that two forms end up in the same class of ECUs has little bearing on those forms’ proximity to each other in the ECUs. It simply indicates that each of these forms appears more frequently in the ECUs of this class than in the ECUs of the other classes, but this does not mean that these words will appear together in the same ECUs, nor that they were ever used in the same sentence or even in the same meaningful utterance. The presence of two significant forms in the same class may be explained by a strong proximity that each form has with a third form, or even with several other forms. These results do not really indicate the extent of the co-occurrences of the two forms in each ECU, and so there is a risk of over-interpretation.

It is within this context of producing factors relating to proximity that Alceste provides a “calculation” of repeated segments. A repeated segment

36. This makes it possible to see how an interviewee’s discourse is deconstructed into the different classes.
37. This is possible if one class is opposed to all of the others in a table.
38. The partitioning of the ECUs remains arbitrary.
is a segment of text consisting of several consecutive forms that appear in the corpus at least twice.

Alceste refines the analysis using a hierarchic ascendant classification (HAC), with the aim of classifying the vocabulary within a single class. The table crosstabulates the ECUs belonging to the class with the root forms found in that class. In this case, the root forms are classified. The closer these forms are situated to each other (within the same ECUs), the more readily they will be grouped together.

2. Application of the Software to a Hip-Hop Corpus

2.1. Our Choices

The twenty interviews analyzed in our study were not initially conducted with computer analysis in mind. They were transcribed after recording, in accordance with general practices: the transcriber provides the context of the interview (including information such as pre-interview arrangements, setting, and other people present) and then transcribes, as faithfully as possible (in other words, without making any “corrections”), the sociologist’s questions and the interviewee’s responses. Other detailed information may be given, for example, relating to the subject during the interview, such as moments of silence, smiles or laughter, a movement of the arm, or relating to external interruption (from the telephone, for example). All of these elements must be included in a discussion of Alceste, as the program is designed to examine repetitions in the discourse or, more specifically, in the corpus.

First of all, we made the choice to eliminate any commentary about the interview situation, as this could not be considered speech (as it was not provided by the interviewer or the interviewee). Even if these elements are significant, they cannot be examined on the same level as the actual words spoken during the interview. The word “laughs,” for example, sometimes appears in the interviews, but it is never actually spoken and so it is problematic to treat it as such (as we have seen, Alceste leaves no choice in this matter).

Although choices relating to corrections were more difficult, we decided to proceed with some of them. First, we homogenized the spelling of certain words among the different interviews (particularly words concerning hip-hop technique). We also inserted underscore symbols in certain composite words. These corrections were made with the sole

39. These words come from English, as in “smurf,” “break,” “battle,” and “freestyle.” This vocabulary does not appear in French dictionaries, which prevents Alceste from recognizing them when analyzing a French corpus as in the present case.
objective of instructing Alceste to consider these expressions as a whole unit, not to be separated.

In terms of corrections, the choice remained of whether to rewrite the corpus in more formal language, particularly where the modification of words containing apostrophes was concerned (such as c’ait, the informal, shortened form of c’était [it was], or p’tit instead of petit [small]). Proceeding with such corrections would certainly increase the number of forms recognized and analyzed by Alceste, thereby most likely increasing the percentage of classified ECUs, but it would also mean losing distinguishing variants. For our present purposes, we ran Alceste without these corrections.

Finally, we chose to remove every utterance attributable to the interviewer.40

2.2. Results

Within the limitations described above, we can interpret the results obtained for our corpus.

We observed that 63% of the ECUs remained in the stable portion, meaning that 37% of the ECUs will not be analyzed. This is not an insignificant loss. A review of the rejected portion shows that some of the abandoned ECUs contained information that was far from secondary to the focus of our research.

Four stable classes were constructed from our data (see fig. 1 below). These classes were not all equal in size; the first two alone contained nearly 76% of the total classified ECUs. Conversely, these two classes contained the fewest significant root forms (the chi-square value exceeded 1541). These two “lexical worlds” appeared more frequently in the stable portion and were, at the same time, focused on certain forms, which resulted in a lower chi-square value for these forms than for the most significant forms in the other classes.42 These two classes were less distinctive than the other two.

40. One test in which the sociologist’s utterances are retained plus other tests “with corrections” are presented in Dalud-Vincent 2010a.
41. In this case, we rejected the null hypothesis if the chi-square value exceeded 3.841 (with a 5% significance level).
42. For class 1, the most significant form is fille (χ² = 78); for class 2, it is jeune (χ² = 125); for class 3, it is y (χ² = 562); and for class 4, it is Käfig (name of dance company) (χ² = 286).
Figure 1: General Results

Class 1
The significant presences43 are as follows: fille/girl (78), break (64), garçon/boy (59), aimer/like (57), sol/floor (51), figure/trick (39), entraîner/to train (36), franchement/frankly (34), battle (31), choré/choreographer45 (30), super/great (28), regarder/to watch (28), trouver/to find (27), force/strength (25), tomas/flare (21), copain/friend (20),

43. Figure and tomas appear exclusively in this class.
44. For simplicity, the root forms are given without suffixes; the chi-square values are given in parentheses. [Translator’s note: the interview corpus is in French, [and so the data displayed here indicates the French term in italics, followed by the English translation. Where no French term appears, either the English term was used in the French text, or the French and English terms are identical.]
45. Familiar, shortened form of the French choréographe.
The language in this class is focused on the technical and athletic aspects of hip-hop dance, and it highlights the similarities and differences between girls and boys. It is concerned with describing how they learn “on the job,” without any real training program. The adjectives and adverbs indicate quite positive judgments, and copain/friend and groupe/group conjure up a sense of sociability.

In this class, there was fewer forms than elsewhere (according to the significant absences) relating to professionalization and training. The terms contemporain/contemporary and classique/classical are also rarely present. Little mention is made of time.

The language conveys a vision that is fairly distant from institutional values, and is instead much closer to the language of the “street” dancer. It has the highest presence in terms of the number of classified ECUs.

Class 3
The significant presences in this class are as follows: y/there (562), contemporain/contemporary (391), z (250), Grenoble (204), p (160), hip-hop (134), technique (119), taire/to silence (102), état/state (99), classique/classical (92), esprit/mind (86), claquette/tap (80), improviser/to improvise (80), tit (74), nant (74), Cunningham (74), Sami (67), danse/dance (66), capoeira (61), f (59), Hachichi (58), voir/to see (57), exactement/exactly (56), jazz (52), Merce (52), Saint-Martin d’Hères (52), f-sait (51), vache/cow (51), Session (51), découvrir/to discover (49), base/basics (46), v (44), nir (44), vice/vice (44), fond/foundation (42).

46. Saint-Martin d’Hères, Merce, Total, concevoir, Los Angeles, Hypnose, vision, versa, New York, ballon, purement, métisse, fusionner, maghrébin, pointe, barrière, sauce, and exemple appear exclusively in this class.
47. In most cases this is p’tit broken into “p” and “tit.”
48. This includes the word tait, from c’était.
49. From maint’nant, for maintenant/now.
50. Choreographer. His first name is Merce, which appears later.
51. Dancer and choreographer. His last name is Hachichi, which appears later.
52. From vach’ment, for vachement, a familiar term for “terribly.”
53. Title of a festival (to be linked with “total” for Total Session.)
54. From “vice versa” (to be linked with verser/to pour).
bottom55 (44), total (44), vision (44), mentir/to lie (44), Hypnose56 (44), concevior/to perceive (44), Los Angeles (44), transmettre/to transmit (43), évoluer/ change (41), chais57 (39), dev (37), ballon/ball (37), verser/to pour (37), New York (37) Allemagne/Germany (36), gestuelle/ gesture (34), africain/African (32), France (31), Bo58 (30), évoluter/to perform (30), attitude (30), associer/to associate59 (30), institution (30), ça/that (29), côté/side (29), sujet/subject (29), métisse/mixed (29), purement/purely (29), enfermé/confined (29), enseigner/to teach (29), États-Unis/United States (28), parler/to speak (26), façon/way (25), monter/to put up (25), manière/manner (25), instant (25), niveau/level (24), ancient/old (23), capable (23), uni60 (22), sauce (22), garde/guard61 (22), chaud/hot (22), pointe (22), exemple/example (22), biennale/biennial (22), cotoyer/to be close with (22), barrière/barrier (22), maghrébin/North African (22), fusionner/to fuse (22), précurseur/precursor (22), fin/end or fine62 (21), connu/known (20), spectacle/performance (20), échange/exchange (19), époque/era (18), problème/problem (18), nu/nude63 (17), chuis64 (17), Sidney (17), genre (17), mondial/worldwide (17), élève/student (16), rythme/rhythm (16), developer/to develop (16), hype (15).

The only notable absence in this group is petit/little (-17). This absence can be linked to “p” and “tit,” which were very present in the same class. The language in this class is probably used, and been faithfully transcribed, in a more “street” way. This is manifest in the “y” and “z” (used in the informal, shortened expression y’z’étaient for ils étaient, or “they were”). Also in this class was taire, whose third-person singular conjugation is tait (homonymous with the informal c’était for c’était, or “it was”). “Nant,” “f,” “v,” “nir,” “chais,” “nu,” and “chuis” are also word fragments that do not appear in class 1.

---

55. General part of the expression à fond l’ballon, meaning “give it everything you’ve got” (to be linked with ballon/ball).
56. Dance company.
57. From je sais, “I know,” or je ne sais, “I don’t know.”
58. Dancer.
59. From “associable.”
60. From États-Unis, or United States.
61. From garder, to keep, or r’garder, for regarder, to watch.
62. From enfin, or finally.
63. From rev’nu (revenu, returned), or dev’nu, (devenu, become).
64. From je suis, I am.
In this class, many names of places, people (dancers and choreographers), and dance companies were mentioned, as well as events such as the Biennale/Biennial (in Lyon, France) and “Total Session.”

The form “hip-hop” (danse/dance appears to a lesser extent in this case) has a very significant presence. Some root forms refer to other types of dance. This class is situated within a more general discussion of the place of hip-hop in a cultural context. Technique and learning are not disregarded—forms such as technique, base/basics, and gestuelle/gesture attest to this. Technique is not described in detail, however, as it is in class 1. Verbs such as improviser/to improvise, voir/to see, découvrir/to discover, transmettre/to transmit, and évoluer/to perform indicate a certain receptivity, dynamism, and positioning with respect to other things. This may also be linked to état/state, esprit/mind, and “vision.”

Class 2
The significant presences are: jeune/young or youth (125), professionnel/professional (85), cours/class or course (69), année/year (57), centre/center (57), demander/to ask (56), action (55), petit/little (54), stage/workshop or internship (51), an/year (48), projet/project (46), formation/training (46), bac/baccalaureate examination (43), vrai/true (43), argent/money (43), âge/age (42), maison/house (42), public/audience (41), heure/time (39), donner/to give (38), animation/activity (38), atelier/workshop (37), manger/to eat (36), intervenir/to intervene (36), école/school (34), vacance/vacation (34), père/father (33), social (32), scolaire/scholastic (31), régulier/regular (31), Danse-Ville-Danse (31), mois/month (27), balle (26), payer/to pay (26), habiter/to live (26), diplôme/degree (26), Saint-Priest (26), conservatoire/conservatory (26), soir/evening (25), semaine/week (25), travail/work (25), mère/mother (24), financier/to finance (24), poser/to ask [a question] or to place [an object] (21), venir/to come (21), Belfort (21), examen/test (21), théâtre (21), Jeunesse et Sport/Youth and Athletics (21), lycée/secondary school (20), cadre/frame (17), semblant/to seem (17), déplacement/movement (17),

65. Bac, animation, scolaire, Danse-ville-Danse, habiter, Saint-Priest, conservatoire, balle, financer, Belfort, Jeunesse et Sports, examen, semblar, pro, and déplacement appear exclusively in this class.
66. From “cent balles,” an expression similar in meaning to “in your dreams.”
67. From professionnel.
First and foremost, this class is about “youth.” It includes forms related to training and preparation for professional life. The issue of money and finances is also very present, as is Time, mentioned in terms such as year, age, and time of day. Certain places and institutions, particularly for young people, also appear. Family relationships are present and cities and events are mentioned. In this class, there is very little reference to technique or to the dance companies of the region.

**Class 4**
The significant presences are as follows: Käfig (286), aujourd’hui/today (266), Mayada (259), compagnie/company (212), pièce/piece (183), Maison de la Danse (113), envie/desire (84), Na70 (82), écrire/to write (77), proposer/to propose (77), programmateur/scheduler (72), soutenu/supported (70), train/train (65), spectacle/show (65), titre/title (64), synopsis (64), effectivement/effectively (64), livre/book (63), administration (63), créeer/to create (61), claquer/slam (57), quitter/to quit (56), nom (54), Mo72 (51), avis/opinion (50), carrière/career (50), rôle (43), Darmet73 (43), Versus74 (43), numéro/number (43), réseau/network (43), épingle/pin (43), sentiment/feeling (43), subvention/subsidy (43), Orphée (42), histoire/story or history (40), directeur/manager (38), festival (37), pont/bridge (36), vendre/to sell (35), affirmer/to affirm (35), collectif/collective (35), programmer/to organize (35), jeu/game (34), politique/politics (33), thématique/thematics (33), Ka76 (32), metteur/director (31), dire/to say (29), merci thank you (29), ponder/produce (29), gueule/face (29), appel/call (28), prendre/to take (29), relation (29), rechercher/next (29), fondateur/founder (29), installer/to establish (29), douloureux/painful (29), confidant/confidant

68. From professeur.
69. Mayada, programmateur, synopsis, titre, claquer, carrière, Darmet, Versus, réseau, sentiment, rôle, épingle, numéro, chuter, pont, douloureux, pondre, merci, fondateur, gueule, recevoir, lire, Traction, construction, planning, article, presse, and étape appear exclusively in this class.
70. Choreographer and dance company director.
71. From en train de, or “in the process of.”
72. Dancer.
73. Director of Maison de la Danse.
74. Performance.
75. From couper les ponts, an expression similar in meaning to “sever all ties.”
76. Dancer.
77. Very familiar.
(28), artiste/artist (27), mode/fashion (26), Accrorap78 (26), parti/position (25), obliger/to oblige (24), diriger/to direct (23), écriture/writing (23), Montpellier (23), page (21), lire/to read (21), tirer/to pull79 (21), étape/step (21), lourd/heavy (21), presse/the press (21), batter/to beat (21), Traction80 (21), intérêt/interest (21), article (21), recevoir/to receive (21), délibéré/deliberate (21), planning (21), valorisé/enhance (21), technician/technician (21), apercevoir/to perceive (21), fonctionner/to function (21), particulier/particular (21), construction (21), nouvelle/news (20), date (19), prise/outlet (19), parcours/path (19), souhait/wish (19), profiter/to take advantage (19), mort/death (17), homme/man (17), éviter/to avoid (17), monter/to put up (17), manager/to manage (17), réponse/response (17), tourner/to turn (17), raison/reason (16), partir/to leave (16), période/period (16), intérieur/interior (16), rencontrer/to meet (16), moment (15), continuer/to continue (15).

Similarly to class 3, this class contains a low proportion of the classified ECUs, but it is highly distinctive. The language in this class focuses on institutions, regional dance companies, performances, dancers, and individuals of note in the field of dance. The organization and creation of performances (professionalization and management) seem to be at the heart of the discussion. Certain forms indicate a certain confidence or experience has been acquired, sometimes going so far as to conjure up a sort of freedom (envie/desire, proposer/to propose, soutenu/supported, créer/create, avis/opinion, relation, souhait/wish) and the taking up of a clear position (claque/slam, quitter/to quit, parti/position, obliger/oblige, batter/beat, éviter/avoid, partir/to leave), but with some pain (lourd/heavy, douloureux/painful).

In contrast to class 1, which focused on “street dancers,” the discussion here is more linked to institutions. This opposition draws attention to the two HDCs, with links between classes 1 and 3, and 2 and 4. This proximity can be seen in the first axis of the factorial correspondence analysis.

Finally, we note that half of the interviewees used a technical, “street dancer” type of language (class 1), and a third referred to training and professionalization (class 2). The other classes contained fewer people (characteristically, at least). One interviewee seems to be highly atypical, as he is an extreme outlier in class 4 with a record-breaking chi-square value.

78. Company.
79. From “tirer son épingle du jeu,” an expression similar in meaning to “play one’s cards right.”
80. From “Traction Avant,” the name of a dance company.
In conclusion, this rather crude interpretation of the data remains consistent with the analysis results of Faure and Garcia. It is subject, however, to many limitations, the main one being that Alceste does not, at any point, incorporate the sociologist’s hypotheses or theoretical viewpoint. The program imposes its own theoretical point of view, for there is certainly a practical and theoretical viewpoint superimposed onto the corpus, as we have shown. Having said that, a few choices do fall to the sociologist, and these are tested elsewhere (Dalud-Vincent 2010a; 2010b).

**References**


81. Due to space restrictions, we will not discuss the other results provided by this program.

82. The only command available, other than the formatting of the corpus, is the display of results and the settings. Is this not “analyzing the corpus”?
— 2001b. Le glossaire de la méthode Alceste. La terminologie utilisée avec des informations techniques. Versailles, France: CNRS-Printemps, Université de Versailles-Saint-Quentin en Yvelines.