Analysing conversational data with computer-aided content analysis: The importance of data partitioning

Dominique Peyrat-Guillard[1], Caroline Lancelot Miltgen[2], Stephanie Welcomer[3]

[1] GRANEM, Université d'Angers, France – dominique.peyrat-guillard@univ-angers.fr
[2] CREM, Université de Rennes, France – caroline.miltgen@orange.fr
[3] University of Maine, USA – welcomer@maine.edu

Abstract

This article highlights the distinctive outcomes generated by different approaches to computer-aided content analysis, and discusses how partition decisions reveal or conceal possible data interpretations. Drawing on data collected from focus groups set up during a European research study, we demonstrate how the chosen encoding technique leads to different views of the same texts, regardless of the software chosen. This analysis produces a user's guide for researchers who need to analyze conversations and concludes with a discussion of the implications for management and organization research.

Keywords: Computer-aided text analysis, computer-aided content analysis, statistical analysis of qualitative data, focus groups, encoding techniques

Funding: This study was funded by the European Commission IPTS (Institute for Prospective Technological Studies) Joint Research Centre under EC JRC Contract IPTS n° 151592-2009 A08-FR.

1. Introduction

Though rich in potential, flowing conversations often offer challenges in analysis. Our primary aim in this study is to show how to analyze conversational data using computer-aided text analysis (CATA) thus illuminating some key data partition decisions the CATA researcher should consider. The article's second contribution is to provide an example of CATA conversational analysis that uses focus groups conducted for a European research study. In this example, we explore the conversation produced during multiple focus groups and compare partitioning by conversation group ("crowded") with the partitioning by individual ("decrowded" or individualized) technique (Bonneau & Dister, 2010 ; Guerrero and al., 2009). We test ways to take the conversation flow into account and show how different encoding techniques might lead to different outcomes in analyses of the same corpus. In addition to these two research objectives, we also aim to demonstrate that the partition choices made by the researcher affect the results of the analysis, regardless of the software chosen. Even if the differences seem relatively finely-grained, they can lead to potentially different empirical results and interpretations of data and consequently different theoretical implications. We provide a detailed examination of computer-aided text analysis of conversations and suggest criteria for selecting the appropriate partitioning approach given the research aim and the particularities of the data.

2. Two CATA approaches and data partitioning

In conducting data analysis with CATA the researcher assigns the level at which the data will be partitioned, dictating how the text will be grouped. For instance, many documents such as

newsletters, circulars, operating procedures, brochures, labor-union pamphlets, motions (Hetzel and al. , 1998), mission statements, corporate annual reports (Mercier, 2002), and press articles (Zueli, 2010) are grouped at the document level because they contain one voice (though they may have multiple authors). In these cases the partition is the document itself. However, the analysis is more complicated when the text is a conversation, such as in individual interviews (Roure & Reinert, 1993), group discussions, group interviews, focus groups (Guerrero et al. 2009), Delphi group interviewing, or reports of meetings. More generally, transcription of spoken material "poses difficult structural problems", as one can read in the Guidelines for Electronic Text Encoding and Interchange. Specialists can refer to the last version of these guidelines (TEI P5, 2014, available online http://www.tei-c.org/), which have been developed and are maintained by the Text Encoding Initiative Consortium (TEI). Our objective in this paper is not interchange of information. Our work mainly targets non-specialists who need to analyze transcribed spoken material and have to choose, for multi-voiced data, between a "decrowded" or a "crowded" approach. A decrowded approach prioritizes on each participant and segments the text to track each individual's statements. In contrast, a crowded approach encompasses the entire interview and thus the conversation flow.

Partition issues are therefore particularly critical in data involving conversations, such as those occurring in focus groups. Focus groups have been increasingly used in social science (Parker & Tritter, 2006) and business marketing (Calder, 1977) research. Focus groups are those groups assembled "for an in-depth exploration of a topic about which little is known" (Stewart & Shamdasani, 1990: 102). A series of open-ended questions are asked and it is hoped that "a kind of momentum is generated which allows underlying opinions, meanings, feelings, attitudes and beliefs to emerge alongside descriptions of individual experiences" (Parker & Tritter: 2006: 2). Focus groups produce data from collective discussions about a defined topic, instead of individual responses to formal questions (Denzin & Lincoln, 2000). Analyzing interactions among the participants (Gersick, 1989) and shared significances thus is particularly important (Banks, 1979). However, the greatest advantage of focus groups—to produce interactions—is often poorly accounted for in literature that uses computer-aided content analysis (Parker & Tritter, 2006). Analysts often focus on individual opinions expressed in group setting, which "is a pity, since the debate flow and block-off processes are usually more relevant" (Mendes de Almeida, 1980: 115). Furthermore, despite the popularity of focus groups as a method of data collection, "there has been relatively little critical discussion of the problematic aspects of conducting focus groups or analyzing the data derived from them" (Parker & Tritter, 2006: 23-24). Therefore, we consider computer-aided analysis of data collected during focus groups, as a basis of a larger critical discussion of computer-aided analysis of data based on conversations.

Researchers using focus groups face issues of generating themes from words grouped by individual and then compared across individuals, versus words used by the group in a complete focus group session –essentially tracking participants' trails of statements through the conversation or tracking the flow of concepts through the group's conversation. In using CATA software, the researcher, of necessity, makes partition choices that generate trade-offs in terms of thematic classification. In the following examples we illustrate how partition issues in two computational content analysis programs, Alceste[1] and WordMapper, yield

[1] A huge proportion of researchers use Alceste for content analysis

distinct results. Notably WordMapper and Alceste represent two different approaches to data analysis, thus offering a means to generalize findings beyond one specific type of software.

3. Method

3.1. Focus group topic

The focus groups we consider in this study discuss citizens' attitudes and behavior toward personal data management and privacy. The objective of the study, which was funded by the European Commission, was to understand European people's use of (or intention to adopt) electronic services (e.g., e-government) and related identity-authentication techniques, as well as their associated motivations, fears, and perceived risks to do so. Interestingly, age seems to affect digital privacy concerns, but empirical findings as yet are mixed. Age appears to dictate how people relate to information technology in general and the Internet in particular, a phenomenon that also influences how concerned they are about their privacy (e.g., Moscardelli & Divine, 2007). A review of recent research on this topic (see Lancelot Miltgen & Peyrat-Guillard 2013) shows that young people can be either more or less preoccupied with privacy than adults, with one study concluding there is no difference (Hoofnagle and al., 2010). Studies focusing on the consequences of young people's privacy concerns on their subsequent behavior similarly have resulted in contradictory results, including both increased protective and risky behavior. These conflicting results regarding the effect of age have prompted calls for cumulative research in this area. To fill this gap, we set up two focus groups in each chosen country, one with young people and the other with adults, to identify possible differences and similarities between these two age categories.

3.2. Sample and data

Two focus groups, with "young people" (15–24 years of age) and "adults" (older than 24 years) respondents, spanned seven EU27 countries (Estonia, France, Germany, Romania, Poland, Greece, and Spain), for a total of 139 participants divided in 14 focus groups, with the same topic guide in each.[2] The interviews were conducted in the country's official language by an academic of a partner University; they were video and audio recorded and then translated into English.[3] Questions and comments from the moderators were removed in the first analysis, as is typical of similar studies (e.g., Guerrero et al., 2009). However, we also tested whether including the comments of the moderators may be of interest as a means to account for different moments of the discourse or conversation. We personally conducted the focus group interviews in France and therefore test our two encoding techniques with this portion of the overall corpus. We checked our results through a comparison with the results from Germany, a country from the same regional block but a different cultural background,[4] as well as a comparison based on the software packages used, to help confirm the generalizability of the results. Specifically, we began with Alceste and then compared our findings with WordMapper, both of which support computer-aided content analysis but with different techniques and hierarchical classification methods.

[2] Available on request.

[3] This translation choice has both advantages and drawbacks, as we discuss in the Conclusion section.

[4] The results for the German corpus are coherent with those we obtained from France; these results are available on request.

3.3. Data analysis techniques

To begin our consideration of focus groups partitioned by individual (decrowded) or the group level (crowded), we compare Alceste's presentation of data for each partition application. We apply both decrowded and crowded encoding techniques to the textual data obtained from the two age-based focus groups from France. The two focus groups' transcriptions are put in the same file to test the influence of age group (young people or adults) on results, this variable being a supplementary element. The format of the French data with the Alceste software, using the decrowded encoding technique, appears in Appendix 1. In the decrowded technique, each participant is the sampling unit of analysis assigned by the researcher (an "initial contextual unit" – ICU) so that the Alceste file combines two focus groups and starts with the discourse of participant 1 (first French participant) and ends with the discourse of the groups' last participant - numbered 20. Participants 1–10 constitute the first focus group (young people under 25 years of age), and participants 11–20 represent the second group (adults older than 24 years). Each of the 20 participants in the first file is represented by demographic characteristics (e.g., gender, age), which were collected through a questionnaire completed by the focus groups participants before the start of the interview. So in this first file, we have partitioned the corpus into 20 Initial Contextual Units (ICUs). With the second encoding technique (see Appendix 1), participants' responses are not partitioned by individual but instead follow the flow of the focus groups' conversation, with the possibility of identifying each participant during the focus group interview, as we have added speakers' names in transcripts at turns of talk (see Krippendorff, 2004: 282). Consequently, in this second file, the corpus is partitioned into 2 focus groups (Youngs and Adults), i.e. 2 ICUs and, in each ICU, the corpus is further partitioned to distinguish turns of talk, with the participant's first name and number. The beginning and end of the corpus is not the same because the first participant saying something during the first focus group was participant number 7. Thus we only have two initial contextual units of analysis at the group level: one for the first focus group (ICU 1, "Young people") and one for the second focus group (ICU 2, "Adults"). As each participant speaks, her or his name and number appears, according to the following code: *NameofParticipant_ParticipantNumber. With this crowded encoding technique, interactions between participants and the flow of ideas are traceable.

4. Results

Even if the number of classes (3 with Alceste Software) is the same with the two encoding techniques, the differences in the content of these classes are important to consider.

4.1. Comparison of the decrowded and crowded encoding techniques

Alceste examines each ICU and breaks them into smaller units ("elementary contextual units" – ECUs). These ECUs are equivalent to a sentence or a paragraph, depending on the length of the corpus and are either ultimately "classified" or left unclassified based on the words in the ECU. Alceste creates a data matrix based on words' presence or absence in the ECUs and performs a Descending Hierarchical Classification (DHC, see (Reinert, 1998 ; Schonhardt-Bailey, 2005)). In our example, Alceste identified 524 elementary contextual units and yielded 435 classified – this 83% classification rate for the entire corpus is considered a highly satisfactory result for individual or group interviews (Reinert, 1998). With the crowded technique, we obtained more ECUs (761) because the turns at talk are taken into account. The percentage of classified ECUs and number of classes was equivalent, though in Figures 1 and 2, the vocabulary of each class and the form of the tree graph differed slightly. These

differences, though sometimes subtle, lead to different interpretations of data and have an impact on the findings. For the decrowded approach (Figure 1), the DHC first separates Class 1 from Classes 2 and 3. Then, it separates Class 2 from Class 3.

The decrowded technique has regrouped, within the same ECU, the following sentences stated by the same person[5] (ECU 387, participant 16, adult, $\chi^2 = 11$):

"But I think that Facebook will hurt the actual generation for work later. They really displayed and this will be prejudicial.

Secondary school pupils don't realize that there can be a potential danger.

It doesn't bother me to be filmed but in my opinion this isn't a way to reduce crimes."

This assignment is potentially problematic because these three sentences were stated at three different moments of the interview. The topic being discussed evolved across these three moments, and other participants provided their opinions in the meantime. The first two sentences refer to the first topic of Class 1 (risks taken by the young generation) but the third sentence refers to the second topic of Class 1 (surveillance cameras). That is why Class 1 is not homogeneous and refers to two different topics. It is thus difficult to interpret this class and to find an appropriate label and the one we have chosen here i.e. "Risks and Dangers & Privacy invasion" reflects this lack of homogeneity.

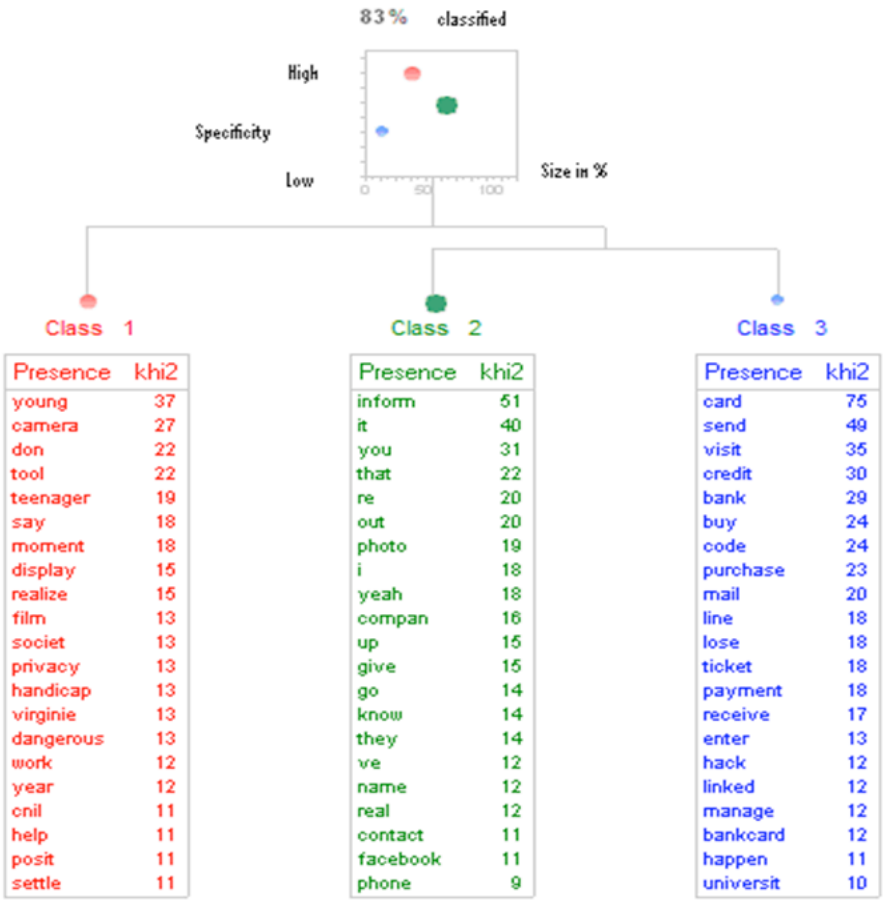

Figure 1. DHC, Alceste Software, Decrowded Encoding Technique - Tree graph

[5] Sentences stated by different persons can never be aggregated in the same ECU.

The crowded encoding technique instead assigns these three moments to three different ECUs: ECU 657, Class 1, $\chi^2 = 9$; ECU 659, Class 1 $\chi^2 = 6$; and ECU 688, Class 3, $\chi^2 = 13$. The first two moments appear in close proximity and the third comment was spoken later in the interview, as we can see through the ECUs' numbers.[6] Furthermore, this last sentence refers to a different topic (i.e., surveillance cameras), so it is classified into another class (3) in the crowded technique. On Figure 2 we can see that, with the crowded technique, the tree graph first separates class 2 from the two other classes and then separates Class 1 from Class 3. Classes 1 and 3 that appear on figure 2 correspond to Class 1 on figure 1. So instead of having a class pertaining to two different topics as we have with the decrowded approach (class 1, figure 1), we obtain two separate classes (class 1 and class 3, figure 2) with the crowded technique. Class 1 and Class 3 of the crowded technique (figure 2) are thus both homogeneous and have been labeled "Risks and dangers" for Class 1 (risks taken by the young generation) and "Privacy invasion" for Class 3 (surveillance cameras to prevent risks). Both refer to sentences stated mainly by adults (χ^2 for adults = 137 for Class 1 and 68 for Class 3). For young people, Class 2 (χ^2 of "Aged 19–24" = 110) of the decrowded technique (figure 1) illustrates their fewer privacy concerns. For example, in ECU 222, Class 2, χ^2 = 10), we found the following statement:

"For me honestly what I would be worried about is my bank details. Apart from that, I really don't mind putting my information on Facebook or whatever, I know I've got nothing to hide."

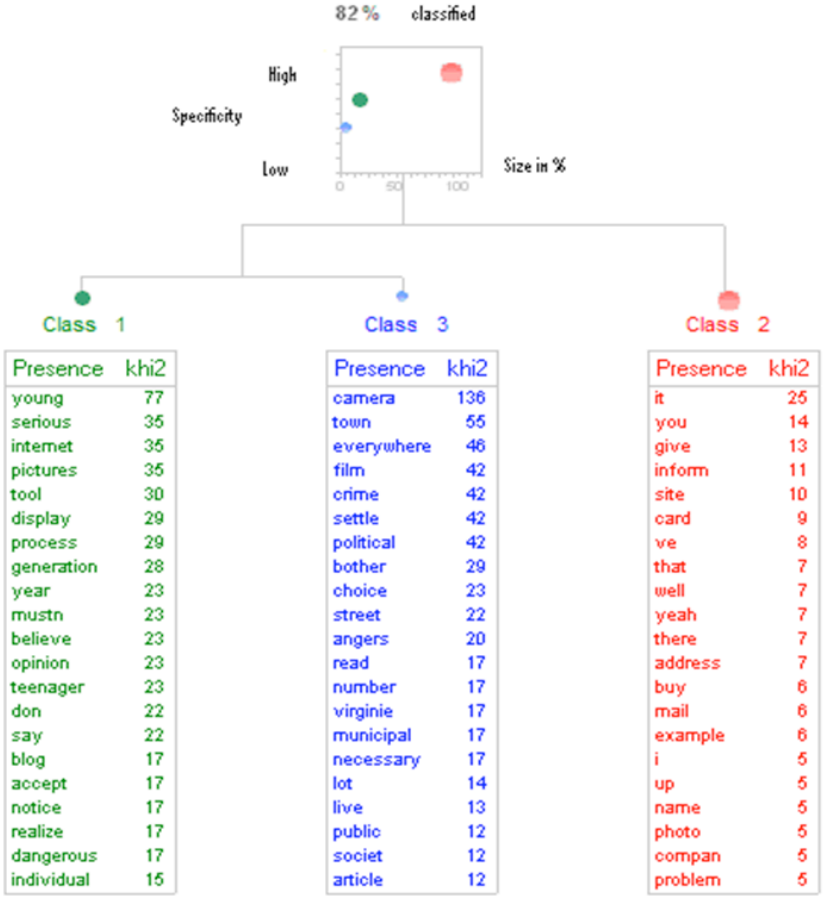

Figure 2. DHC, Alceste Software, Crowded Encoding Technique - Tree graph

[6] ECUs numbers respect the order of the transcription.

Class 2 on Figure 1 is quite homogeneous and has been labeled "Data disclosure on SNS". However, Class 3 in this decrowded technique does not offer a clear pattern of themes or participants. It is specific to neither young people nor adults (χ^2 of "Aged 15–18" = 3; χ^2 of "Aged more than 61" = 7), and the vocabulary pertains to different contexts. The most representative ECUs differ, which makes it difficult to interpret the meaning. For example, the following two ECUs are characteristic statements from Class 3 (figure 1):

"When I go into my account, I have to enter a code, for example C3 and each time I make a different payment it's a different code" (Young participant aged between 19 and 24).

"Well I'm Maurice, I'm retired. It didn't happen to me. I only visit scientific sites. No purchase on internet, no sites such as Facebook or anything like that" (Adult participant aged more than 61.)

Of key importance is that the classes are more homogeneous with the crowded technique thus making it easier to interpret and discern potential constructs of interest. For example, Class 2 in the crowded technique reveals the links between data disclosure, data use and responsibility. This class refers to sentences pronounced mainly by young people (χ^2 for young = 230). The topic of responsibility was not so obvious with the decrowded technique. Yet responsibility is an important topic and in the crowded approach's results, a clear generational divide appears, such that for privacy issues, younger respondents are more responsible and confident than older adults. This finding is in line with some results in prior literature showing that young people are more self-confident Internet users, as 'digital natives' (Baumann, 2010). However, our results also contradict conventional 'scare-mongering' about young people's online lives (Herring, 2008), which depict them as reckless and ignorant. Instead, young people take a greater degree of personal responsibility. This more nuanced picture of young people portrays them as active agents, engaging with privacy in different ways than older generations, but not lacking in concern about access to and control of their personal data.

4.2. Generalization: A test with another software package (WordMapper)

We applied these two encoding techniques to the same corpus (without moderator questions), using a different software package to test the robustness of the results and/or determine if we could find any differences. We chose WordMapper, a text-mining software package that can perform a hierarchical classification but does so in ascending order. Instead of separating the words used in different contexts, as Alceste does (i.e., a principle of DHC), WordMapper regroups the words used in the same contexts to build clusters (i.e., a principle of AHC). It is thus interesting to compare the results of the classifications performed in descending and ascending order on the same corpus. To prepare the corpus, we separated its different parts: A row beginning and ending with the characters "[" and "]" plays the same role as the row beginning with the character "*" in Alceste. The number of clusters is higher than with Alceste, 12 for the decrowded technique and 11 for the crowded one. This is logical for an ascending classification compared to a descending one.

More time is needed to code the text with the crowded technique (row with "[….]" for each intervention during the focus group), but, as with Alceste software, the clusters were more homogeneous than those obtained with the decrowded technique and thus easier to interpret. As we noted previously, the decrowded technique cannot distinguish different moments in the focus group interviews. Thus it provides excerpts such as the passage below, which corresponds to the last four sentences pronounced by Isabelle, participant 19 in the adult group. She begins by referring to surveillance cameras, and ends with an unrelated reference

to using magnetic swipe cards. This extract appears in the cluster named "people touch"[7] by the WordMapper software. This cluster is not homogeneous in the case of the decrowded technique because it refers to two different topics: surveillance cameras and data disclosure and use:

"*We are watched but we don't know where they are. If it would be necessary to have cameras, it would be necessary to know where they are precisely. Regarding our liberty we're not going to live in this street. It's like the purchase you pay with your credit card."*

These sentences were classified by WordMapper in the same cluster, because they were pronounced by the same person not necessarily at the same time but in this order. Clearly the first three sentences appeared much earlier than the last sentence, and many other persons expressed their views in the meantime, which shifted the topic. An excerpt from the crowded corpus related to these sentences instead revealed:

[0019 ; Isabelle ; FGA[8]] "*We are watched but we don't know where they are. If it would be necessary to have cameras, it would be necessary to know where they are precisely. Regarding our liberty we're not going to live in this street".*

… (discourse of twelve other different participants)…

[0016 ; Xavier ; FGA] "*So what is the purpose we can wonder? Concerning the setting up of the tramway in the town there will be a pass too and it will be magnetic, without any contact where people will be able to be followed in the same way*."

[0019 ; Isabelle ; FGA] "*It's like the purchase you pay with your credit card."*

The topic here is no longer surveillance cameras in the streets; it has shifted to being tracked through the use of a magnetic card or a credit card. However, with the decrowded technique, the last sentence of Isabelle was classified in the cluster "people touch" whereas the intervention of Xavier just before was classified in another cluster ("address e-mail") relating to data use. As this example shows, in the decrowded technique, two words pertaining to different topics can be classified within the same cluster. The crowded technique prevents such problems, leading to more homogeneous clusters that are easier to label and interpret. With the crowded technique, Isabelle's last sentence, given as example before, was classified by WordMapper in the cluster "money time" related to data use. The sentence stated just before by Xavier was classified in another cluster named "address e-mail" related also to data use and very close to the cluster "money time" on the MDS graph which confirms their topic proximity. Furthermore, with Alceste software, the crowded technique makes salient the topic of responsibility, in relation to data disclosure and use on SNS.

As a further illustration, we compared the classifications of certain words according to the encoding technique and software used. It shows that, with the decrowded technique, two words pertaining to different contexts (for example young and camera for Alceste and camera and card for WordMapper) are classified in the same class or cluster by both software packages. Thus, the corresponding classes or clusters are not homogeneous. On the contrary, with the crowded technique, the words pertaining to different contexts belong to different classes or clusters and all the classes or clusters are homogeneous. Consequently, the crowded technique highlights the differences between young and adults and the attendant importance of a topic like responsibility, in relation to data disclosure and use.

[7] The names of the clusters are defined by WordMapper, according to word's frequency and co-occurrences.

[8] FGA = Focus Group Adults

5. Discussion

The encoding technique has a clear impact on the data analysis results. The differences noticed in this work are subtle but important. Each technique offers its own advantages and drawbacks. With the decrowded technique, the number of partitions of the corpus is determined by the number of respondents so that the time required to encode the file is minimal. The crowded encoding technique involves more intensive scanning over the entire corpus so it is a more time-consuming exercise. However, it also gives a very different reading of the corpus as distinct moments of the conversation and the interactions between the participants are taken into account. Such distinctions are not possible with the decrowded technique so that the interpretability of the hierarchical classifications (descending or ascending) are clearer with the crowded technique and reveal some aspects that pass unnoticed with the decrowded technique. With the crowded approach the classes or clusters are more homogeneous and it is easier to label them. Moreover, the crowded approach was better suited here to reveal a reversed "privacy paradox" for young people: lower privacy concerns combined with a greater use of protection strategies. Studies that indicate greater privacy concerns among young people refer to their use of protection behaviours; we suggest instead that low privacy concerns can combine with high protection strategies, which clarifies the consequences of young people's privacy concerns as linked to protective behaviours instead of risky ones. By analyzing these conversations with two content analysis programs corresponding to two different approaches (a dictionary approach for Alceste and a statistical association approach for WordMapper) we aim to demonstrate that these partition trade-offs are not software-specific.

In summary, the crowded technique offers more advantages than drawbacks when the goal is to analyze the emergence of themes in a conversation. It is a less applicable partition choice if there are many short answers with poor content appearing as separate parts of the corpus. These short answers cannot be classified by a content analysis software package. The advantages of the crowded approach may apply to the interventions of the moderator or interviewer, but with a caveat. Incorporating moderators into the analysis appears to influence the results; it seems however necessary to make the distinction between focus groups and group interviews when considering this question because the dynamic differs in each case. That is, "In group interviews the researcher adopts an 'investigative' role: asking questions, controlling the dynamics of group discussion, often engaging in dialogue with specific participants," whereas in a focus group, the researcher plays the role of a "facilitator/moderator of group discussion between participants, not between her/himself and the participants" (Parker & Tritter, 2006:25-26). For focus groups the interventions can be short and of "poor" content, which can reduce the percentage of classified answers. It thus seems more appropriate to account for the moderator's questions in group interview settings. However, in each case, the best solution is to compare the results obtained with and without the interventions, to gain complementary views of the same corpus. In our case, the percentage of classified answers was still greater than 50% when we included the interventions of the moderator.

6. Conclusion

The differences we have noted reflect our investigation of one part of the corpus, which constitutes a limitation of this study. Investigating the two encoding techniques across the entire corpus (139 participants, seven countries) would offer a stronger test of our results. In

this study, as seven countries were involved, a common language was necessary for analyses. Another limitation thus pertains to our translation of the focus group discussions into English, which prevents any consideration of specific national vocabularies. This gap poses a challenge to CATA for studies that involve multicultural issues and/or multiple countries and languages. Is it better to translate all the discourses into one language (which permits a direct comparison but ignores languages' specificities), or is it preferable to retain the original languages? Further research should investigate this question to define the conditions in which each solution is optimal. Despite these limitations, our work demonstrates the influence of encoding techniques on the results of computer-aided content analyses and offers a user's guide to researchers willing to analyze any type of conversation. As a first generalization test we examined the two encoding techniques with the German corpus: the outcomes are very similar, implying a solid basis for the possible generalization of our results. In particular, the results from the German corpus confirm that short answers are better accounted for with the decrowded technique[9]. Our findings also suggest that it is important to analyze a corpus using different software packages, not only content analysis programs as it was the case in this study but also text analysis programs because "they have the potential to operate in concert" (Wolfe et al., 1993: 645). This combination would help to obtain an interesting, complementary view of the same text and thereby improve the quality of the interpretation.

References

Banks J.A. (1979). Sociological theories, methods and research techniques. a personal viewpoint. *Sociological Review*, 27 (3), 75-84.

Baumann M (2010). Pew report: digital natives get personal. *Information Today*, 27 (10).

Bonneau J. and Dister, A. (2010). *Logométrie et modélisation des interactions discursives. L'exemple des entretiens semi-directifs*. Journées Internationales d'Analyse Statistique de Données Textuelles (JADT). International conference on textual data analysis, 9-11 June, Rome.

Calder B. J. (1977). Focus groups and the nature of qualitative marketing research, *Journal of Marketing Research, 14(August),* 353-364.

Denzin N.K. and Lincoln Y.S., eds. (2000). *Handbook of qualitative research* (2d ed.). Thousand Oaks, CA: Sage.

Gersick C. J. G. (1989). Marking time: predictable transitions in task groups. *Academy of Management Journal*, 32 (2), 274-309.

Guerrero L., Dolors Guardia M., Xicola J., Verbeke W., Vanhonacker F., Zakowska-Biemans S., Sajdakowska M., Sulmont-Rossé C., Issanchou S., Contel M., Scalvedi M.L., Signe Granli B. and Hersleth M. (2009). Consumer-driven definition of traditional food products and innovation in traditional foods. A qualitative cross-cultural study. *Appetite*, 52, 345-354.

[9] In the case of the German corpus, the results obtained with the crowded technique were not very good for the first test: only 36% of classified ECUs instead of 69% with the decrowded encoding technique. These results are due to many short answers in the German focus groups. These short answers have a poor content and thus cannot be classified by the software. To check this, we have deleted all the short answers of this kind in the German corpus and have reanalyzed the corpus. For this second test, the percentage of classified units is much higher, at 65% and is very close to the percentage obtained with the decrowded technique. The number of classes is also the same as the one with the decrowded technique whereas it was different with the corpus including short answers. These results are coherent with those obtained when we take into account the interventions of the moderator.

Herring SC (2008). Questioning the generational divide: Technological exoticism and adult constructions of online youth identity. In Youth, Identity, and Digital Media, 71-92, Massachusetts Institute of Technology.

Hetzel A-M., Lefèvre J., Mouriaux R. and Tournier M. (1998). *Le syndicalisme à mots découverts – Dictionnaire des fréquences (1971-1990)*. Paris: Syllepse.

Hoofnagle C., King J., Li S. and Turow J. (2010). How different are young adults from older adults when it comes to information privacy attitudes and policies, available at: http://ssrn.com/abstract=1589864.

Krippendorff K. (2004). *Content Analysis – An Introduction to Its Methodoloy*. 2nd ed. Thousand Oaks: Sage.

Lancelot Miltgen C. and Peyrat-Guillard D. (2013). Cultural and generational influences on privacy concerns: a qualitative study in seven European countries, *European Journal of Information Systems,* Advance online publication 30 July2013; doi: 10.1057/ejis.2013.17

Mendes de Almeida P. (1980). A review of group discussion methodology. *European Research*, (May), 114-120.

Mercier S. (2002). Une typologie de la formalisation de l'éthique en entreprise: l'analyse de contenu de 50 documents. *Revue de Gestion des Ressources Humaines*, 43, 34-49.

Moscardelli DM. and Divine R. (2007). Adolescents' concern for privacy when using the Internet: an empirical analysis of predictors and relationships with privacy-protecting behaviors. *Family and Consumer Sciences Research Journal*, 35(3), 232-252.

Parker A. and Tritter J. (2006). Focus group method and methodology: current practice and recent debate. *International Journal of Research & Method in Education*, 29 (1), 23-37.

Reinert M. (1998). *Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste*. Communication aux JADT, Journées Internationales d'Analyse Statistique de Données Textuelles/International Conference on textual data analysis. Nice, France.

Roure H. and Reinert M. (1993). *Analyse d'un entretien à l'aide d'une méthode d'analyse lexicale*. Journées Internationales d'Analyse Statistique de Données Textuelles (JADT). International Conference on textual data analysis*,* 21-22 October, Montpellier, France.

Schonhardt-Bailey C. (2005). Measuring ideas more effectively : An analysis of Bush and Kerry's national security speeches [online]. London : LSE Research Online. http://eprints.lse.ac.uk/archive/00000862

Stewart D. W. and Shamdasani P. N. (1990). *Focus groups: Theory and practice.* Newbury Park, CA: Sage.

Zueli C. (2010). *Using computer-assisted text analysis to identify media reported events*. International conference on textual data analysis (JADT)*,* 9-11 June, Rome.

Appendix 1 : Encoding techniques applied

Original corpus without moderator's comments (two focus groups in the same file)	**Beginning of the corpus, i.e. beginning of the focus group with youngs:** Christophe – participant number 7 *For example you've got those MSN sites where you can, if you want to, fill in your profile, your name, your surname, address…. Personally, I didn't fill that part in for example…* **Later, during the focus group with youngs:** Pierre – participant number 1 *It depends on the way you use it as well, it depends on the use. For myself I don't hesitate to put heaps of information on Facebook. That doesn't bother me. Anyone can get in contact with me, there's my phone numbers, addresses, it really doesn't bother me…* **Beginning of the focus group with adults:** Maurice – participant number 11 *Well I'm Maurice, I'm retired…* **Later, during the focus group with adults:** Virginie – participant number 20 *I read an article recently I don't know in which country it is, in any country except France. And they proved that there was less crime before the settlement of the cameras than afterwards. That is to say that they did not avoid anything at all.* **End of the corpus, i.e. end of the focus group with adults:** Jean-Sébastien – participant number 12 *It's like computing there is always a parade.*
Encoded corpus with the first encoding technique: decrowded, i.e. partition by individual. Alceste software	**Beginning of the corpus – discourse of the participant number 1 during focus group with youngs** 0001 *Country_France *Gender_Male *Age_19-24 *It depends on the way you use it as well, it depends on the use. For myself I don't hesitate to put heaps of information on Facebook. That doesn't bother me. Anyone can get in contact with me, there's my phone numbers, addresses, it really doesn't bother me.…* **End of the corpus – discourse of the participant number 20 during the focus group with adults:** 0020 *Country_France *Gender_Female *Age_25-44 *… I read an article recently I don't know in which country it is, in any country except France. And they proved that there was less crime before the settlement of the cameras than afterwards. That is to say that they did not avoid anything at all.*
Encoded corpus with the first encoding technique: decrowded, i.e. partition by individual. WordMapper software	**Beginning of the corpus – discourse of the participant number 1 during focus group with youngs** [0001 ; Country_France ; Gender_Male ; Age_19-24] *It depends on the way you use it as well, it depends on the use. For myself I don't hesitate to put heaps of information on Facebook. That doesn't bother me. Anyone can get in contact with me, there's my phone numbers, addresses, it really doesn't bother me.…* **End of the corpus – discourse of the participant number 20 during the focus group with adults:** [0020 ; Country_France ; Gender_Female ; Age_25-44] *… I read an article recently I don't know in which country it is, in any country except France. And they proved that there was less crime before the settlement of the cameras than afterwards. That is to say that they did not avoid anything at all.*
Encoded corpus with the second encoding technique: crowded, i.e. partition by focus group and by turns of talk. Alceste software	**Beginning of the corpus – focus group with youngs (coded FG_Youngs) :** 0001 *Country_France *FG_Youngs -*Christophe_7: *For example you've got those MSN sites where you can, if you want to, fill in your profile, your name, your surname, address…. Personally, I didn't fill that part in for example. …* **Later in the corpus:** -*Pierre_1: *It depends on the way you use it as well, it depends on the use. For myself I don't hesitate to put heaps of information on Facebook. That doesn't bother me. Anyone can get in contact with me, there's my phone numbers, addresses, it really doesn't bother me.…* **Beginning of the focus group with adults (coded FG_Adults):** 0002 *Country_France *FG_Adults -*Maurice_11: *Well I'm Maurice, I'm retired….* **End of the corpus, i.e. end of the focus group with adults:** -*JeanSebastien_12: It's like computing there is always a parade.
Encoded corpus with the second encoding technique: crowded, i.e. partition by focus group and by turns of talk. WordMapper software	**Beginning of the corpus – Focus Group with Youngs (coded FGY) :** [0007 ; Christophe ; FGY] *For example you've got those MSN sites where you can, if you want to, fill in your profile, your name, your surname, address…. Personally, I didn't fill that part in for example. …* **Later in the corpus:** [0001 ; Pierre ; FGY] *It depends on the way you use it as well, it depends on the use. For myself I don't hesitate to put heaps of information on Facebook. That doesn't bother me. Anyone can get in contact with me, there's my phone numbers, addresses, it really doesn't bother me.…* **Beginning of the Focus Group with Adults (coded FGA):** [0011 ; Maurice ; FGA] *Well I'm Maurice, I'm retired….* **End of the corpus, i.e. end of the focus group with adults:** [0012 ; JeanSebastien ; FGA] It's like computing there is always a parade.