

## Chapter 11

# A Brief History of Software Resources for Qualitative Analysis

### 11.1. Introduction

This chapter is intended as a resource for qualitative analysis of corpora of texts in human and social sciences. We will concentrate on the computerized analysis of text corpora<sup>1</sup> by obscuring the tools to analyze documents that are not strictly textual. These can, for example, be objects confided by informants, memories that the observer retains from his/her contact with the field or recordings (audio, photographic or video) of interactions or settings.

Tools to analyze these non-exclusively textual materials (as *Aquad*, *Anvil*, *Atlas*, *Porphyr*<sup>2</sup>, *Transana*, *Transcriber* and *Videograph*) are therefore not analyzed here. This methodological work leads to the opening of some black boxes on textual analysis. Such a clarification will bring to mind the circumstances in which different techniques<sup>3</sup> develop and shows that many tools under different names, resume very similar features. For this reason, I prefer to identify families of features rather than families of tools (such typologies are available in [JEN 96, KLE 01, POP 97, WEI 95]).

---

Chapter written by Christophe LEJEUNE.

<sup>1</sup> A corpus is a set of documents. For this chapter, it is a set of texts.

<sup>2</sup> See Chapter 1.

<sup>3</sup> Here I focus on the methodological dimension of these tools, the issue of interpretation being addressed in Part Five of this treatise.

In this chapter, I would like to show those scared of this technique that these tools are close to their daily work without software. To the enthusiasts, this chapter points out that none of the proposed techniques is magic, or a Pandora's Box or intelligent, because they are no more than tools. Finally, those of my readers who are already using some of the mentioned tools will perhaps interpret this chapter as a strong overview. Indeed, like any list, this overview covers tools developed in very different worlds. I hope it will encourage dialogs among users from different communities.

### 11.2. Which tool for which analysis?

As the anthropology of science has brilliantly shown for the other disciplines, scientists transform materials observed into inscriptions. From translation to translation, introductory measurements are moved (mobile) from the phenomenal field to the laboratory [LAT 88]. The inscription (on a support) ensures their stability during this transfer. Given that their ability to represent the phenomenon is not altered, they are said immutable. The converging movement from the phenomenal field towards the laboratory is therefore attached a movement that is moving away from it: the researcher mobilizes the inscriptions in the assembly of argumentative features that are articles [CAL 86]. It is the quality (both mobile and immutable) of these inscriptions that ensures fidelity to the starting observations. By vocation, these new inscriptions start a centrifugal displacement in relation with the laboratory.

In social sciences, the series of transformations into “immutable mobile” starts with the collection (and registration) of the testimony of the informant. This is then taken to the laboratory; scientists conduct other transformations (translations) to it that produce new inscriptions. Among them, the transcription of interviews prototypically occupies the introductory position of translation, making the following transformations possible. Even when no other tool is then used, this operation is indeed necessary to the quotation (without which the empirical basis of sociological arguments is largely weakened). Once the interviews are transcribed, the researcher often opts for the analysis of these intermediary inscriptions.

The following sections review the devices likely to accompany this task. The features presented will be:

- the felt-tip (section 11.2.1);
- text processing (section 11.2.2);
- the operating system – and, in particular the regular expressions – (section 11.2.3);

- the cotext – in particular, the concordances – (section 11.2.4);
- the co-occurrence (section 11.2.5);
- Benzécri analysis of data (section 11.2.6);
- text segments (section 11.2.7); and
- dictionaries (section 11.2.8).

### 11.2.1. *The felt-tip*

The minimum equipment the qualitative researcher needs in order to analyze his/her textual material is the felt-tip (or, if preferred, the pen or pencil). Without this starting point being the opportunity to discuss the link between science and writing [GOO 77, SER 02], it nevertheless recalls that empirical sources manipulated by the researcher are material. Faced with empirical elements that are transcriptions of interviews, the analyst proceeds to the identification of themes that seem relevant to him/her.

The felt-tip here is the minimum technology required to allow this annotation. Since the present chapter focuses on computer tools, this borderline case stands as non-included endpoint (which means that this chapter does not further study work with felt-tip itself)<sup>4</sup>.

### 11.2.2. *Text processing*

Some researchers propose to take advantage of the familiarity and spread of text processing and use them to analyze the corpus of text [LAP 04, MOR 91]. This specific use of an office tool consists of delimiting segments of text and associating them with labels chosen by the researcher<sup>5</sup>. Two mainly features are used for this purpose: invisible marks and tables.

Initially designed as adjuvant to writing, invisible marks allow the user to comment on his/her text (without it appearing in printing); they are thus particularly taken advantage of in the case of collective writing. Used as part of a qualitative analysis, these revision marks can annotate the text, transposing – on screen – the use of the felt-tip. Depending on the strategy of the researcher, these annotations may go from a simple account of sayings to an interpretation.

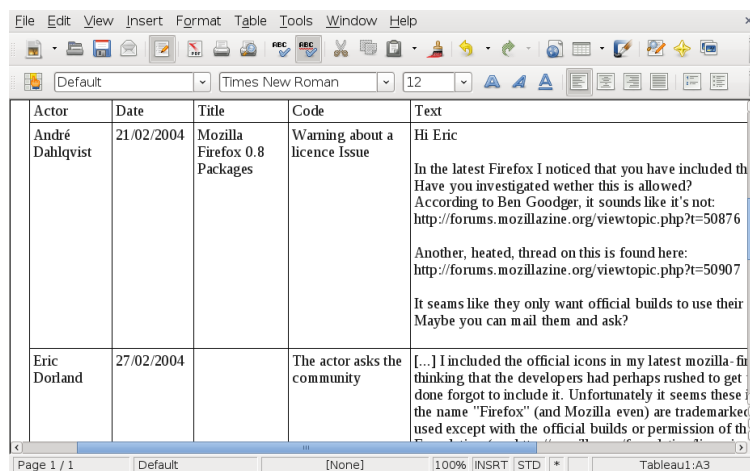
---

<sup>4</sup> In section 11.2.7 I show that some software design makes reference to the technology of the felt-tip.

<sup>5</sup> Close to work “with felt-tip”, this deterritorialization of text processing is consistent with tools widely-spread among people (that I present in section 11.2.7).

Second, as shown in Figure 11.1, tables can also be used as a counterpart: annotations then no longer lay in the invisible marks but in a dedicated column (which also resembles paperwork but, this time, comments made in the margin)<sup>6</sup>.

The use of text processing to analyze qualitative material is attested and thus possible. However, experience shows that this diversion is not the most appropriate tool<sup>7</sup>.



Actor	Date	Title	Code	Text
André Dahlqvist	21/02/2004	Mozilla Firefox 0.8 Packages	Warning about a licence Issue	Hi Eric  In the latest Firefox I noticed that you have included the Have you investigated whether this is allowed? According to Ben Goodger, it sounds like it's not: <a href="http://forums.mozillazine.org/viewtopic.php?t=50876">http://forums.mozillazine.org/viewtopic.php?t=50876</a>  Another, heated, thread on this is found here: <a href="http://forums.mozillazine.org/viewtopic.php?t=50907">http://forums.mozillazine.org/viewtopic.php?t=50907</a>  It seems like they only want official builds to use their Maybe you can mail them and ask?
Eric Dorland	27/02/2004		The actor asks the community	[...] I included the official icons in my latest mozilla-firefox thinking that the developers had perhaps rushed to get done forgot to include it. Unfortunately it seems these icons the name "Firefox" (and Mozilla even) are trademarked and used except with the official builds or permission of the

**Figure 11.1.** Word processing (Open Office). (The codes of the researcher appear in the fourth column)

### 11.2.3. The operating system

In practice, researchers who use text processing as analysis tool take advantage of the availability of other word processor features. The exploration and coding of their textual material involves localization of words (or groups of words or parts of words) within the text. As such, these search operations on strings of characters belong to the know-how of each person rather than to the scientific analysis of qualitative material.

<sup>6</sup> The illustration presented here uses the free word-processing program *Open Office*, <http://www.openoffice.org/>, accessed February 8, 2010.

<sup>7</sup> On large corpora, tabular files become less easy to handle.

The search for alphanumerical segments is not limited to typographical strings. It includes famous truncations – that allow us, for example, to locate all the conjugated forms of the same verb (or, more generally, all the inflections<sup>8</sup> of the same lexeme [LYO 95]) as well as operators, such as the logical `or`. The range of patterns that can be gathered is thus infinite.

Such localization of generic patterns (“pattern matching”) uses what computer scientists call “regular expressions” [FRI 06]. These have been available on all operating systems since the birth of microcomputers. They were born from the scientific study of neurons. In the 1940s, the neurophysiologist Warren McCulloch browsed various disciplines to understand how thought develops in mankind. His meeting with the talented logician Walter Pitts lead to an article [MCC 43] that tried to model the nervous system. Propositional calculation was applied to these neuronal machines (communicating between themselves by electrical pulse). For McCulloch, this work fits into a constant questioning of the (material) basis of the human mind and the willingness to give to this question a scientific (rather than metaphysical) answer, thanks to experimental psychology helped by biology. These latter disciplines will ignore this work<sup>9</sup>.

The works of McCulloch were then resumed in mathematics. Stephen Kleene attached the concept of finite automation to them and included them in his algebra of regular sets<sup>10</sup>. After several developments in mathematics, the regular expressions were gradually introduced into the concerns of computer scientists. The co-developer of *Unix* introduced them in the 1960s both in scientific literature [THO 68] and in a series of computer applications (first the editor *qed*, then *ed*, that popularize them). In these editors, the following command should have been entered to execute a connection of patterns:

```
g/Regular Expression/p
```

---

8 Among the morphological transformations, linguists distinguish the inflection of the same lexeme in different word-forms (of which the German versions and the conjugation of verbs in English are examples) of the derivation of a new lexeme (to go from a noun like “territory” to a verb like “territorialize”)

9 The aura surrounding the 1943 article - considered a precursor of neural networks and cognitive science [AND 92, DUP 09] - calls for some restrictions: with such a force of attraction, it can be quoted without necessarily being the most relevant reference for the foundation of pattern matching. Its history shows nevertheless how various concerns (scientific, philosophical and epistemological) were able to converge in the development of a generic tool.

10 In French, “regular” is sometimes translated as “rational”, which led some French researchers [SIL 00] to speak of rational expressions [FRI 06].

This function gave its name to the application *grep* (for *global regular expression print*).

The power of regular expressions then increased [FRI 06]. In the late 1980s, the language Perl<sup>11</sup> played a decisive role in their spread. However, Larry Wall, who designed of this programming language, is a linguist [SCH 08]. Regular expressions can thus boast about their interdisciplinary origins.

Today, they are part of the analysis of text corpora. This section thus suggests that the researcher can be satisfied with using the operating system to analyze his/her data, without having recourse to other more specific tools than the search functions<sup>12</sup>.

#### 11.2.4. *The cotext*

By leaving the use of conventional office tools, the present picture offers a class of more specific tools. Close to previous search features, these tools are used to identify words, expressions or patterns. Their specific contribution lies in the surrounding words of the element sought. Exhibiting the sentence or paragraph in which the target appears allows the researcher to relate each occurrence to a context of apparition (what linguists call “immediate topological environment” or cotext [KER 02, MAI 98, WIL 01]).

This mode of visualization of the passages coming before and after each of the occurrences is called “concordances” [LEB 98] or “index of keywords in context” [WEIT 95]. Concordances<sup>13</sup> are typically presented with vertical alignment of the target (or pole-shape), so that different cotexts can easily be sorted, brought closer to one another and compared [PIN 06] (see Figure 11.2).

---

11 <http://www.perl.org/>.

12 Jocelyne Le Ber thus deterritorialised the *grep*, *freq* and *diff* commands of the operating system in order to analyze literary works, in particular *Antigone* by Jean Cocteau [LEB 06].

13 *AntConc* by Laurence Anthony (<http://www.antlab.sci.waseda.ac.jp/>) ; *Glossanet* by Cédric Fairon (<http://glossa.fltr.ucl.ac.be/>) [FAI 06].

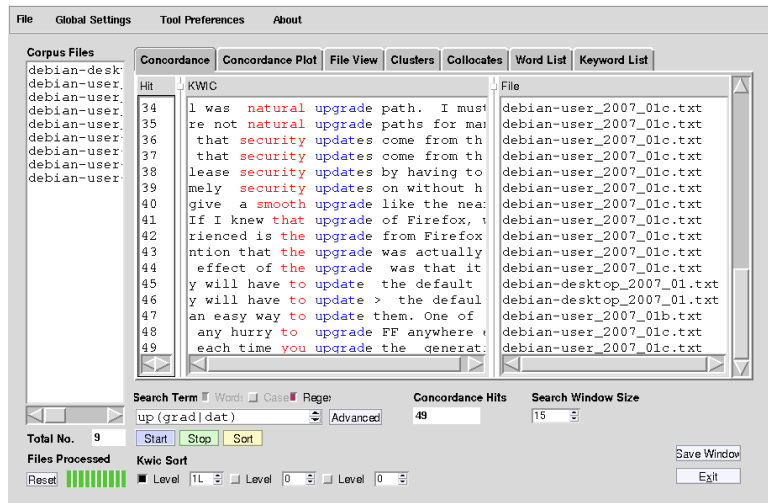


Figure 11.2. A concordancer (AntConc)

The history of indexes dates back to the beginning of our era. The first indexes organized in alphabetical order dated from the 4<sup>th</sup> Century AD. The Greeks compiled (non-alphabetical) indexes of geographical names in the 6<sup>th</sup> or 7<sup>th</sup> Century but the invention of concordances and indexes of keywords in context fits into the tradition of exegetes of sacred texts. A division of biblical texts into chapters (called “capitulation”) was tested in the 4<sup>th</sup> Century [MEY 89]. The text of the Carolingian *New Testament* contains, in the margin, a capitulation as well as references to similar passages in other Gospels. The proliferation of books and the reduction of volume sizes reduced the room available for such information. In parallel, exegetes produced specific tools to ensure this intertextuality. In the 10<sup>th</sup> Century, the Masoretes – the Jewish “masters of tradition” – created alphabetical lists of words accompanied by their immediate cotext<sup>14</sup>. They therefore foreshadowed (or invented) the keywords in context [WEIN 04]. At the dawn of the 13<sup>th</sup> Century, Etienne Langton introduced chaptering in the Latin bible<sup>15</sup>. Between 1238 and 1240, the Dominican friars of the Saint Jacques convent in Paris, under the direction of Hugues of St Cher, made an alphabetical list of concordances involving these new references. These include the chapter number and a letter (A to G), indicating the position of the word in the chapter. The cotextual environment being unavailable

14 These lists were not indexed, given the absence of numbering of chapters or verses.

15 The verses were inserted, much later, by Robert Estienne in 1551.

in this edition, the system in question prefigures for its part the indexes of keywords *out of* context. In subsequent editions, the framing proposition was added as a third element. It is precisely the debates between Christians and Jews that supported the development, by Rabbi Nathan Mardochee, of a Hebrew concordance in 1523, which was the first to be printed [SEK 95].

In the 1930s, the scientific study of the cotext rose in the United States. The context is no longer religious, but political: the sponsors of these studies were dealing with American Indian languages that were as numerous as they were poorly known. Inspired by naturalistic approaches, structural linguistics then developed and introduced the notion of context : according to this notion, each language element is defined by its textual environment. The set of various environments of an element is known as its distribution [HAR 64]. Although distinct from concordances, distributional analysis marks the integration of cotext into scientific tools.

The next innovation in concordances lay in automation. The first automated generations of concordances were proposed by late 1950s<sup>16</sup>. In 1959, Hans Peter Luhn proposed a setup of concordances focusing on the desired form, which he qualified under the acronym *KeyWords In Context* or KWIC [LUH 60, LUH 66]. Some years later (in 1963) the first indexes of keywords designed according to this principle appeared (the target words were from titles of 10 years of publication of the *Association of Computer Machinery* [YOU 63]).

After centuries of existence, the construction of concordances is now automated. The concordancer is thus the tool with the oldest tradition. Its interdisciplinary nature is comparable to regular expressions. It is even currently being extended beyond the study of textual material, with specific applications in genetics and chemistry<sup>17</sup>. The concordances are used to support many inferences of corpus exploration. Their graphical properties make them particularly appropriate to discovering the recurrence of phrases, idioms or expressions (consisting of several simple words).

### 11.2.5. The co-occurrence

Co-occurrence is a variation of the previous cotext-based tools. Although different algorithms exist in this field, the principle is still based on the same logic: identifying the “proximity”<sup>18</sup> of two terms. This proximity is neither semantic, nor

---

16 In 1951, Roberto Busa, author of the afterword of this treatise, using tabulating tools made a concordance of four poems of Thomas d’Aquin.

17 In genetics, concordancers help to locate genes sequence alignment (named “synteny”)<sup>18</sup>

18 One also speaks of the (lexical) attractin (or repulsion) of two terms [HEI 98].

syntactic or pragmatic. As with cotextual tools, it refers to a topological dimension. The two terms are even closer than the fact they are separated by a few characters. This proximity is measured along the syntagmatic<sup>19</sup> axis. To go from this measurement to co-occurrence, we aggregate the proximities of each appearance (or occurrence) of the two terms. Contrary to the study of environments that, through the operating concept of distribution, find a justification in American structuralist linguistics, the proximity of co-occurrence is the subject of relatively few linguistic studies in the strict sense. Attested works are those of Maurice Tournier. His design of proximity is based on the psychological discoveries linked to Pavlovian conditioning and epistemologically accompanied the *Lexico* software by André Salem.

Co-occurrence is therefore a feature closely linked to textual analysis tools. Algorithms vary according to how they measure proximity. Some actually take the number of characters into account; others are based on a count of words separating the two terms. Some consider that proximity can be measured throughout the text; others limit the relevance to one chapter, one paragraph, one sentence or one proposition (these different units of context being mostly defined according to typographical criteria). In contrast to concordance, co-occurrence most often does not take the order of words into account; we speak of pairs of co-occurrences or of the network of words associated with a pole. Some software (such as *Tropes*<sup>20</sup> [MAR 98] and *Weblex*<sup>21</sup> [HEI 04]) nevertheless offer features that distinguish association according to the order in which words appear; we therefore speak of *ordered* pairs of co-occurrences.

In contrast to concordance, co-occurrence is not attached to a typical layout of results. It can be presented either as lists or graphs. The graphs are often not oriented [OSG 59, TEI 91]. The use of oriented graphs is nevertheless attested when the order of words is taken into account. Just like concordances, co-occurrences are used to make inferences. In a logic close to automatic categorization, some tools use this measurement to build aggregates of elements that are strongly linked. These aggregates (or clusters) offer a synthetic view of the corpus and are subject to interpretation by the analyst. In sociology, such topics maps are mobilized by the anthropology of sciences [LEJ 04], in particular within *Candide* [TEI 95] which is itself based on developments of *Leximappe* [CAL 91, LAW 88, VAN 92]<sup>22</sup>.

---

19 Linguists [SAU 83] distinguish the horizontal (syntagmatic) axis of sequence of words in a sentence of the vertical (paradigmatic) axis of combinations of words that “go together” (for example, words that have the same sound, signification or distribution).

20 <http://www.acetic.fr/>, accessed February 8, 2010.

21 <http://weblex.ens-lsh.fr/>, accessed February 8, 2010.

22 These tools are based on research on information systems [CHA 88, MIC 88].

### 11.2.6. Data analysis

In France, the works of the mathematician Jean-Pierre Benzecri gave birth to a series of techniques such as the principal component analysis and factor analysis. Their use by Pierre Bourdieu deeply influenced French sociology. For the author of *The Distinction*, the factorial design has become a system of representation of the social space as a force field [BOU 87]. Two (orthogonal) axes cut through this space and trace four quadrants opening the now-famous combinatorial distribution of cultural and economical capitals. Besides the stroke of genius (proceeding from the comparison of the table and theory of social classes), the encounter of the author and data analysis introduces a graphical inscription of what gives a field of positions relative to one another to sociology. This relational design, dear to Bourdieu, is embodied in a two-dimensional space where there are both individuals and variables, as well as social classes and practices [BOU 98]. Bourdieu's analyses deal with figures and, in general, the data analysis belongs to the arsenal of quantitative methods. This tool was, however, originally developed to be used in linguistics. It is therefore not surprising that it has inspired the development of software commonly used as qualitative tools. It is for this reason, and with respect to the related sociological field, that these tools are mentioned here, although these programs are very close to statistics.

In France, one of the sociological<sup>23</sup> tools most able to take advantage of Benzecri's teachings is *Alceste*, developed by Max Reinert<sup>24</sup>. Even closer to automats than the felt-tip, the heart of *Alceste* lies in the descending hierarchical classification of lemmatized forms<sup>25</sup> of full words<sup>26</sup> of the analyzed corpus. This leads to a series of formally constructed classes. The ascending analysis modules (to highlight the most typical words for each class) were then mobilized and modules of correspondence factor analysis provide the graphical representations.

Besides the factorial plans, the results are expressed in the typical form of oriented graphs called dendograms. This tree-like setup provides the researcher with an illustration of nested aggregates [KRI 04]. As in the case of co-occurrences, these may serve as basis for inferences of the analyst. Max Reinert, who designed *Alceste*,

---

23 It is in the language sciences, more than sociology, that the most orthodox applications of correspondence factor analyses are found, like the *Data and Text Mining* (DTM) tools developed as a result of *SPAD-T* by Ludovic Lebart.

24 *Alceste* is distributed by *Image*.

25 Lemmatization consists of bringing different inflected or derived forms to a common root. This operation responds to the morphological phenomena mentioned previously.

26 Automated approaches sometimes exclude a list of words from their procedures. These discarded words consist of articles, prepositions, pronouns and conjunctions or, more simply, of the most frequent words of the corpus. They are called "empty words" or "stopwords".

maintains that the interest of these classes is essentially exploratory and heuristic. In so doing, he insists on the necessary complementarity of a deep knowledge of the corpus and computer tools (which help to make assumptions rather than to instrument the administration of evidence).

### 11.2.7. *Segments of text*

Co-occurrence and data analysis demonstrated automatic ways of labeling and segmenting texts corpora. The manual annotation of segments of texts adopts a totally different strategy. Widespread among researchers, tools based on the coding of segments of text bear the name of *Computer-Assisted Qualitative Data Analysis Software* (CAQDAS)<sup>27</sup>. Often claiming relation to the grounded theory [COR 07], CAQDAS advocates that the researcher immerse him- or herself in the corpus to be analyzed<sup>28</sup>.

Texts belonging to the corpus are read by the researcher. He/she focuses on the passages he/she wishes to: this annotation is done by selection (nowadays, usually with the mouse) and association with a label chosen by the analyst (Figure 11.3). This way of proceeding is very similar, both in ergonomics and philosophy, to the use of the felt-tip. Concerning gesture, the selection recalls the highlighter and labeling, the annotations in the margin. Numerous software programs show a space to the right or left of the text that reproduce the layout of a sheet of paper<sup>29</sup>. In the analysis strategy, these tools keep the coding for the researcher. This way of proceeding is certainly more refined than the automation described in the previously. However, it requires a demanding work, that is laborious when following an amendment of the analysis framework it involves reverting to the whole of the corpus. According to the testimony of its practitioners, its limits are a tendency to focus on coding at the expense of analysis, interpretation and theorizing. Some researchers consider this limitation to thematic statements as a pledge of scientificity; on this point they become comparable to radical ethnomethodology that prohibits any interpretation [GAR 02].

---

27 The use of word processing, presented in section 11.2.2, is a particular case.

28 *WeftQDA* by Alex Fenton (<http://www.pressure.to/qda/>); and *TamsAnalyser* by Matthew Weinstein (<http://tamsys.sourceforge.net/>), accessed February 8, 2010.

29 *NVivo* by Lyn and Tom Richards (<http://www.qsrinternational.com/>) and *MaxQDA* by Udo Kuckartz (<http://www.maxqda.com/>), accessed February 8, 2010.

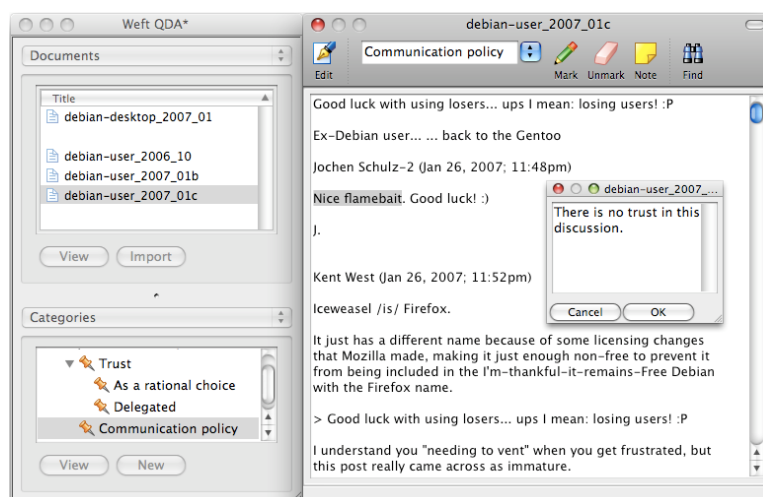


Figure 11.3. Segments of text highlighted in the text (WeftQDA)

If the path of the material is similar to reading on paper, these programs offer advantages linked to digitizing the corpus. It thus becomes easy to locate a passage, through both its content and the labeling that was assigned to it<sup>30</sup>. In the same way, limitations inherent to paper no longer apply; the researcher can thus annotate at leisure a passage that particularly inspired him/her, even if this note is long. As I have mentioned, these tools refer to the analysis “by hand” and are thus part of the techniques of content analysis. As a precursor event to these techniques, specialists have identified the controversy that accompanied the publishing of 90 hymns in Sweden in the 1640s [KRI 04]. Although with the endorsement of the Swedish censor, the “songs of Zion” bothered the Lutheran organization. The content of the collection was the subject of controversy between supporters and critics: for each theme, the frequency and treatment in both the incriminated hymns and classic sources were evaluated. Operated in a contradictory way (by the different protagonists of the controversy), this confrontation is regarded as a forerunner of content analysis.

Techniques of content analysis appeared in the United States at the end of the 19<sup>th</sup> Century in the quantitative, diachronic and comparative study of mass communication, in particular the press. One of the first investigations of this type

30 Such features make use of the tools presented in Paragraph 11.2.3.

was conducted on a corpus covering editions of the *New York Times* over more than 10 years. The magnitude of the subjects examined was then measured in inches (length of articles) [BER 52]. The author of this study lamented on the tendency of newspapers to give an exaggerated importance to short news items and to sensational articles at the expense of substantive articles on politics, literature and religion.

Political sciences then took over these techniques, in particular to study the propaganda broadcast during the two World Wars of the first half of the 20<sup>th</sup> Century. In doing so, the tool was sharpened in order to meet the criteria of a scientific discipline. It is from this period that the flourishing studies of political speeches are dated. It is also during this period that the first automatic systems to systematize coding operations appeared.

From the beginning of computing, researchers in human sciences have developed tools of this type [STO 66]. Of course, they were helped in this endeavor by pioneering engineers in computer science (the very ones that I have shown that forged the regular expressions in particular)<sup>31</sup>. CAQDAS is therefore not very young! Moreover, most of the issues related to qualitative analysis had already been formulated by this period: from the selection criteria of relevant passages and sharing them among analysts, to the necessity of interpreting coding and the validity of these interpretations, through the congruence of analysis with the content of texts.

As I have stated in this section, the central virtue of these segments of text is the appropriateness between coding and the fine-tuned meaning that the interpreter is able to extract from an intimate knowledge of the corpus. The investment required for this return is exhaustive, thorough, even (often) repeated reading. Segments of text tools are thus time-consuming. At the opposite end of the spectrum, automated tools – such as co-occurrence or Benzécri statistics – propose to save time on this substantial work. Such automation however sacrifices some finesse, since the subtleties of the participants' expression and the context of each statement are often lost. There is an intermediate way between reading (and annotation) of the whole corpus by the researcher and the transfer of this crucial task to a machine: this solution is dictionaries.

---

<sup>31</sup> Developed in the 1990s, the *General Inquirer* is still active today (<http://www.wjh.harvard.edu/~inquirer/>, accessed February 8, 2010). Its designer, Philip Stone, died on January 31, 2006

### 11.2.8. Dictionaries

Dictionaries allow automation of coding while not necessarily sacrificing the finesse (and indexicality) of the common meaning studied. The use of dictionaries is based on the fact that there is a stable meaning on which we can lean. These vocabularies may include words or phrases depending on their grammatical category, their (almost) synonymy, or their relevance to the theory of the analyst. Sometimes these lists operate a selection of units on which the analysis is carried out (this is the case of lists of “stopwords” or when only one grammatical category – most often that of nouns – is taken into account).

In a predictable way, tools using grammatical categories are mainly developed within the language of sciences<sup>32</sup>. Those that include synonyms, argumentative records or logics of actions tend to be found in the sciences of culture (such as sociology, history or anthropology). These categories then compose an analysis framework. The question of their relevance to the corpus is variable depending on whether these directories are provided as they are in the tool<sup>33</sup> or are built according to the idiomatic reality of the ongoing research<sup>34</sup>.

### 11.3. Conclusion: taking advantage of software

In social sciences, the features reviewed in this chapter are rarely mobilized independently of each other. It is most often by combining them that software can effectively assist the researcher. For instance:

- the Provalis Research Suite (by Normand Péladeau) provides Benzécri's analyses, concordances, dictionaries and segments of text;
- T-Lab (by Franco Lancia) combines co-occurrences and Benzécri's analyses.
- Sophisticated segments of text – such as MaxQDA – often propose statistical features (or, at least, allow us to export intermediate results to statistical software);
- CAQDAS such as AtlasTI even includes dictionaries [LEW 07].

Inferences therefore proceed by crossing, comparing and cross-checking.

---

32 For example in linguistic engineering, *Unitex* by Sébastien Paumier (<http://www-igm.univ-mlv.fr/~unitex/>) or *Nooj* by Max Silberstein (<http://www.nooj4nlp.net/>); in linguistics of the corpus, *Xaira* by Lou Burnard (<http://www.xaira.org>) and, in statistical analysis of textual data, *Hyperbase* categorized version by Etienne Brunet or *Weblex* by Serge Heiden. Websites accessed February 8, 2010.

33 This former option is consistent with the methods of content analysis [STO 66].

34 This latter case is relevant for interpretative and grounded approaches; for this reason, I propose to speak of registers [LEJ 08].

At the end of this range of possibilities, I hope I have answered the readers' questions or have at least dispelled the shadow hovering around these mysterious software programs in sociological analysis<sup>35</sup>. Whatever the strategy that is chosen, I hope I have shown that there is no technique that ensures the scientificity or originality of research. Whether or not it is computerized, the method requires discipline and, in all cases, the quality of interpretation always comes to the scientist<sup>36</sup>.

In qualitative sociology, the asset of analysis software lies ultimately in the facilitation offered for exchange and discussion among researchers<sup>37</sup>.

Rather than presenting the tool, like a shield against criticism, it is important to open the black boxes and share experiences.<sup>38</sup>

#### 11.4. Bibliography

- [AND 92] ANDLER D. (ED.), *Introduction aux Sciences Cognitives*, Gallimard, Paris, 1992.
- [BER 52] BERELSON B., *Content Analysis in Communication Research*, The Free Press, Glencoe, 1952.
- [BOU 87] BOURDIEU P., *Distinction: A Social Critique of the Judgement of Taste*, Harvard University Press, Cambridge, 1987.
- [BOU 98] BOURDIEU P., *Practical Reason: On the Theory of Action*, Stanford University Press, Stanford, 1998.
- [CAL 86] CALLON M., "Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay", in J. Law (ed.), *Power, Action and Belief: A New Sociology of Knowledge*, p. 196-223, 1986.
- [CAL 91] CALLON M., COURTIAL J.-P., TURNER W.A., BAUIN S., "From translations to problematic networks: An introduction to co-word analysis", *Information on Social Sciences*, vol. 22, no. 2, p. 191-235, 1991.

---

35 This chapter is an invitation to use the available tools: not only domestic software (presented in sections 11.2.2 and 11.2.3), but also free software, whose (spreading and modification) logic is congruent with the scientific mind.

36 Even when we use computers, the construction of an interpretation that is beyond the thematic analysis remains a real challenge. It may only be identified by avoiding the excess theory that forgets (or overwrites) the ground, and naïve empiricism, which neglects interpretative work.

37 Part Five of this treaty continues this discussion. See also [DEM 06].

38 I thank Aurélien Benel, Alex Fenton and Raphael Leplae for their advice. I dedicate this chapter to the memory of my professor of methodology and director of my doctorate at the Institute of Human and Social Sciences at the University of Liege, René Doutrelepon, who died on the April 1, 2005.

- [CHA 88] CHARTRON G., Analyse des corpus de données textuelles, sondage de flux d'informations, doctorate thesis, University of Paris VII, Paris, 1988.
- [COR 07] Corbin, J., Strauss, A., *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, Sage, Thousand Oaks, 2007.
- [DEM 06] DEMAZIÈRE D., BROSSAUD C., TRABAL P., VAN METER K.M., *Analyses Textuelles en Sociologie*, PUR, Rennes, 2006.
- [DUP 09] DUPUY J.-P., *On the Origins of Cognitive Science: The Mechanization of the Mind*, The MIT Press, 2009.
- [FAI 06] FAIRON C., SINGLER J., "I'm like, "hey, it works!": Using Glossanet to find attestations of the quotative (be) like in English- language newspapers", in A. Renouf and A. Kehoe (eds.), *The Changing Face of Corpus Linguistics*, Rodopi, Amsterdam/New York, p. 325-336, 2006.
- [FRI 06] FRIEDL J., *Mastering Regular Expressions*, O'Reilly, Sebastopol, 2006.
- [GAR 02] GARFINKEL H., *Ethnomethodology's Program: Working Out Durkheim's Aphorism*, Rowman and Littlefield, Boston, 2002.
- [GOO 77] GOODY J., *The Domestication of the Savage Mind*, Cambridge University Press, Cambridge, 1977.
- [HAR 64] HARRIS Z.S., "Distributional structure", in J.A. Fodor and J.J. Katz (eds.), *The Structure of Language. Readings in the Philosophy of Language*, Prentice-Hall, New Jersey, p. 33-49, 1964.
- [HEI 98] HEIDEN S., LAFON P., "Cooccurrences. La CFDT de 1973 à 1992", *Des mots en Liberté, Mélanges Maurice Tournier*, vol. 1, p. 65-83, 1998.
- [HEI 04] HEIDEN S., "Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex", PURNELLE G., FAIRON C., DISTER A., (eds), *Le Pouvoir des Mots. Actes des Journées Internationales d'Analyse Statistique des Données Textuelles*, p. 577-588, Louvain, 2004.
- [JEN 96] JENNY J., "Analyse de contenu et de discours dans la recherche sociologique française : pratiques micro-informatiques actuelles et potentielles", *Current Sociology*, vol. 44, no. 3, p. 279-290, 1996.
- [KER 02] KERBRAT-ORECCHIONI C., "Contexte", CHARAUDEAU P., MAINGUENEAU D., (eds.), *Dictionnaire d'Analyse du Discours*, p. 134-136, Seuil, Paris, 2002.
- [KLE 01] KLEIN H., "Overview of text analysis software", *Bulletin of Sociological Methodology*, vol. 70, p. 53-66, 2001.
- [KRI 04] KRIPPENDORFF K., *Content Analysis. An Introduction to Its Methodology*, Sage, Thousand Oaks, 2004.
- [LAP 04] LA PELLE N., "Simplifying qualitative data analysis using general purpose software tools", *Field Methods*, vol. 16, no. 1, p. 85-108, 2004.

- [LAT 88] LATOUR B., *Science in Action: How to Follow Scientists and Engineers through Society*, Harvard University Press, Cambridge, 1988.
- [LAW 88] LAW J., BAUIN S., COURTIAL J.-P., WHITTAKER J., "Policy and the mapping of scientific change: a co-word analysis of research into environmental acidification", *Scientometrics*, vol. 14, p. 251-264, 1988.
- [LEBA 98] LEBART L., SALEM A., BERRY, L., *Exploring Textual Data*. Kluwer, Dordrecht, 1998.
- [LEBE 06] LE BER J., "L'adaptation comme contraction. Une analyse informatique d'Antigone", Duteil-Mougel C., Foulquié B.,(eds), *Corpus en Lettres et Sciences Sociales. Des Documents Numériques à l'Interprétation. Actes du Colloque International d'Albi "Langages et Signification"*, p. 257-268, 2006.
- [LEJ 04] LEJEUNE C., "Représentations des réseaux de mots associés", Purnelle, G., Fairon, C., Dister, A. (eds.), *Le Poids des mots. Actes des Journées Internationales d'Analyse Statistique des Données Textuelles (JADT)*, p. 726-736, 2004.
- [LEJ 08] LEJEUNE C., Au fil de l'interprétation. L'apport des registres aux logiciels d'analyse qualitative, *Swiss Journal of Sociology*, vol. 34, no. 3, p. 593-603, 2008.
- [LEW 07] LEWINS, A., SILVER, C., *Using Software in Qualitative Research. A Step-by-Step Guide*, Sage, London, 2007.
- [LUH 60] LUHN H.P., "Keyword-in-Context Index for Technical Literature", *American Documentation*, vol. 11, n° 4, p. 288-295, 1960.
- [LUH 66] LUHN H.P., "Keyword-in-Context Index for technical literature (KWIC Index)", in D.G. Hays (ed.), *Readings in Automatic Language Processing*, Elsevier, New York, p. 159-167, 1966.
- [LYO 95] LYONS J., *Linguistic Semantics: An Introduction*, Cambridge University Press, Paris, 1995.
- [MAI 98] MAINGUENEAU D., *Analyser les Textes de Communication*, Dunod, Paris, 1998.
- [MAR 98] MARCHAND P., *L'Analyse du Discours Assistée par Ordinateur*, Armand Colin, Paris, 1998.
- [MCC 43] MCCULLOCH W., PITTS W., "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Math. Biophysics*, vol. 5, p. 115-133, 1943.
- [MEY 89] MEYNET R., *L'Analyse rhétorique. Une nouvelle Méthode pour Comprendre la Bible. Textes Fondateurs et Exposé Systématique*, Cerf, Paris, 1989.
- [MIC 88] MICHELET B., L'analyse des associations, Ph.D. thesis, University Paris VII, Paris, 1988.
- [MOR 91] MORSE J., "Analysing unstructured interactive interviews using the Macintosh computer", *Qualitative Health Research*, vol. 1, no. 1, p. 117-122, 1991.

- [OSG 59] OSGOOD C., "The representational model and relevant research methods", in I. De Sola Pool (ed.), *Trends in Content Analysis*, University of Illinois Press, Urbana, p. 33-88, 1959.
- [PIN 06] PINCEMIN B., ISSAC F., CHANOVE M., MATHIEU-COLAS M., "Concordanciers: Thème et variations", in J.-M. Viprey, A. Lelu, C. Condé and M. Silberztein (eds.), *Actes des Journées Internationales d'Analyse Statistique des Données Textuelles*, Besançon, Presses Universitaires de Franche-Comté, vol. 2, p. 773-784, 2006.
- [POP 97] POPPING R., "Computer programs for the analysis of texts and transcripts", in C.W. Roberts (ed.), *Text Analysis for the Social Sciences. Methods for Drawing Statistical Inferences From Texts and Transcripts*, Lawrence Erlbaum, New Jersey, p. 209-221, 1997.
- [SAU 83] DE SAUSSURE F., *Course in General Linguistics*, Open Court, London, 1983.
- [SCH 08] SCHWARTZ R., PHOENIX T., FOY D., *Learning Perl*, O'Reilly, Sebastopol, 2008.
- [SEK 95] SÉKHRAOUI M., Concordances: Histoire, méthodes et pratique. PhD thesis, University of la Sorbonne nouvelle Paris 3 and École normale supérieure de Fontenay St-Cloud, Paris, 1995.
- [SER 02] SERRES M., *Origins of Geometry*, Clinamen Press, Manchester, 2002.
- [SIL 00] SILBERZTEIN M., INTEX, User Manual, 2000, available at <http://intex.univ-fcomte.fr/>, accessed February 8, 2010.
- [STO 66] STONE P.J., DUNPHY D.C., SMITH M.S., OGILVIE D.M., *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge, United States, 1966.
- [TEI 95] TEIL G., LATOUR B., "The Hume machine: can association networks do more than formal rules?", *Stanford Humanities Review*, vol. 4, no. 2, p. 47-55, 1995.
- [THO 68] THOMPSON K., "Programming techniques: Regular expression search algorithm", *Communications of the Association of Computer Machinery*, vol. 11, no. 6, p. 419-422, ACM Press, New York, 1968.
- [VAN 92] VAN METER K.M., TURNER W.A., "A cognitive map of sociological AIDS research", *Current Sociology*, vol. 40, no. 3, p. 123-134, 1992.
- [WEIN 04] WEINBERG B.H., "Predecessors of Scientific Indexing Structures in the domain of religion", *Second Conference on the History and Heritage of Scientific and Technical Information Systems*, p. 126-134, 2004.
- [WEIT 95] WEITZMAN E., MILES M., *A Software Source Book. Computer Programs for Qualitative Data Analysis*, Sage, London/Thousand Oaks/New Delhi, 1995.
- [WIL 01] WILMET M., "L'architectonique du "conditionnel"", *Recherches linguistiques*, vol. 25, p. 21-44, 2001.
- [YOU 63] YODEN W.W., "Index of the Journal of the Association of Computer Machinery", vol. 1-10 (1954-1963), *Journal of the Association of Computer Machinery*, vol. 10, no. 4, p. 583-646, 1963.