Aldo Benini

Text Analysis under Time Pressure

Tools for humanitarian and development workers

Young men in Dili, East Timor, are selling apples and mobile phone cards on the sidewalk. Texts have been printed on T-shirts, apple boxes, and phone cards. None has been produced locally although youth gang membership may influence the choice of idols and thereby of images, logos and texts. Some kind of text analysis arguably happens when buyers copy the credit codes from the cards to their phones. The author, new to this brand, rubbed too fast, deleting a digit. Attempts to infer it from an adjacent card in the perforated sheet predictably failed; the company safeguarded itself against such clever analysis. Rubbing it carefully, even under time pressure, is the way to go.

July 2009 / March 2010

This paper appears in a series of occasional technical notes available at www.aldo-benini.org . Other titles include:

> *Efficient linking of lists in humanitarian data management* (2008)
>
> *The Wealth of the Poor. Simplifying living standards measurements with Rasch scales?* (2007)
>
> *"Runs of relief" - A data management technique for humanitarian logistics analysts* (2007)

This site offers also published papers or links to the journals in which they appeared.

Suggested citation:

Benini, Aldo (2010). *Text Analysis under Time Pressure. Tools for humanitarian and development workers*. Washington, DC [Version 1st March 2010]

Contents

Figures

Tables

Summary

Persons analyzing text documents commonly mitigate time pressure through greater

- selectivity
- substitution
- reliance on computers

or any combination of those. They may bias attention to a segment of the texts to consider. They may substitute for their own reading and analysis the opinions of experts and stakeholders familiar with the texts or their objects. They may use computer programs that assist reading, comprehension, analysis and reporting, beyond the universally used applications.

This paper presents three tools suitable to assist text analysis work that humanitarian and development professionals do under time pressure. These tools are meant to be of help in two basic situations:

- The analysis stays within classic interpretive (i.e., non-statistical) approaches, focused on the intent of the document authors, the texts' internal logic, and the relevant professional and audience contexts. The analyst is helped by rapid navigation and comparison inside and between texts. For a typical situation, think of the briefing papers that an evaluation team receives, some together with the terms of reference, some at the project site.

- Second, by choice or sheer necessity, interpretive approaches are supplemented or, in the extreme, replaced by statistical operations that return distributions and correlations of text elements that carry meaning, notably words and terms. Use of such results may range from navigation to exploration to testing of hypothesis. For example, a federated international NGO may produce major policy documents centrally. Over time, these are reflected in local policy adaptation and implementation documents created in the participating member organizations.

The three tools are aligned with these analysis traditions:

- TextSTAT is a concordance, word frequency and document navigation tool, created by an external researcher. It appeals to the interpretive researcher who finds navigating through hardcopies or in word processors slow.

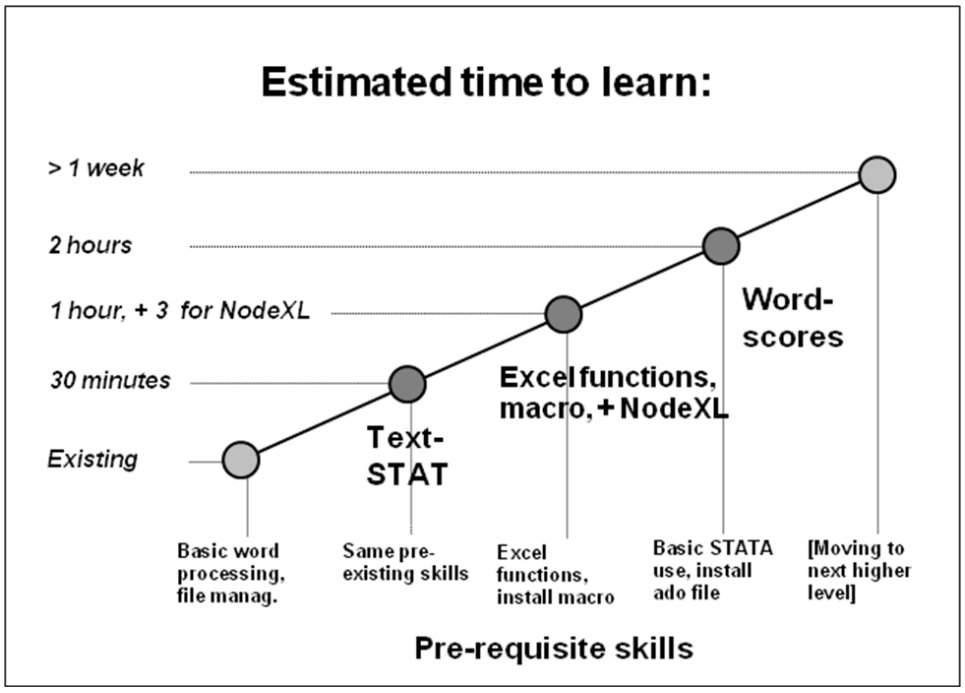
- The popular spreadsheet application MS Excel can be put to uses that straddle both traditions, particularly when the number of distinct texts is significant. I wrote a composite function to flag terms in text stored *inside* the worksheet as well as a macro that extracts term frequencies from a set of *external* Word or .txt documents. I also offer a macro that computes a *term association* matrix, which can be used to graph out relationships among concepts.

There has been a significant change from Excel 2003 to Excel 2007. Excel 2003 was limited to 256 columns per sheet, and text in a cell was limited to 1,024 displayed characters[1]. In the new version, the number of columns is practically unlimited for most purposes (it is over 16,000), and a cell can hold *and display* over 32,000 characters. This opens greater possibilities for analyzing text stored in the spreadsheet itself. In addition, a new tool written for Excel 2007 (NodeXL) affords novices comfortable and integrated access to network graphing. These two developments have prompted a minor revision of this paper (March 2010).

- Wordscores is a collection of routines written for the statistical application STATA. It creates word frequency tables extremely fast, and with the option of an important text preprocessing operation (stemming) that facilitates subsequent statistical analysis.

TextSTAT is freeware. The Excel tools and Wordscores are freely available to users of Excel and STATA. The three tools (plus NodeXL) can be adopted individually, or as steps on a learning curve. The learning investment, given modest skills in the underlying applications, is slight, in keeping with the idea that intending users work under time pressure.

Figure 1: Tools on the text analysis learning curve [incl. NodeXL]



I discuss also the wisdom, or not, of moving up on the learning curve to take advantage of more advanced tools of text analytics and related qualitative research – tools that

[1] The cell could hold 32,767 characters and all of them could be displayed in the formula bar.

require substantially greater learning and, for some, financial investments. I am skeptical of their benefits for most users in this audience. To all practitioners, even those unconcerned with text analysis, I recommend TextSTAT as a free, self-contained productivity tool. Others who wish to explore analysis forms exploiting word (or term) frequencies may find use for the Excel and STATA tools.

> *"In terms of bytes, written words are insignificant, amounting to less than 0.1% of the total [data]. However, the amount of reading people do .. has almost tripled since 1980, thanks all that text on the Internet.*
>
> *The technologies described in this report .. exist to make data more digestible for humans"*
> (The Economist 2010: 5, 17)

This paper remains focused on productivity and on a small number of tools that hopefully enhance it. In various sidebars, I highlight connections to the conceptual hinterland of text analysis, but I do not provide theoretical discussions of meaning structures or of the functions of text documents in the humanitarian and development worlds. These tools float in a stream of fast changing applications as well as methodological and policy fashions. Consider this itself a text permanently in the works.

Acknowledgements

TextSTAT is a freeware tool created by Matthias Hüning at the Dutch Linguistics Institute, Free University of Berlin. It can be downloaded at http://www.niederlandistik.fu-berlin.de/textstat/software-en.html .

Ray Tweedale wrote a macro for wildcard searches from Word to Excel. I used some of its object declarations and the looping structure in my word frequency macro.

Wordscores is a collection of STATA routines written by Ken Benoit, Michael Laver and Will Lowe, freely available from their Web site http://wordscores.com/ . I thank Will Lowe for additional comments. NodeXL is a free Excel 2007 add-in, written by Marc Smith and others, and available at http://www.codeplex.com/NodeXL.

For demonstration material, I used the Oxfam International 2007 Annual Report, the Lutheran World Federation (LWF) information service articles published in the first semester 2008, as well as UN Office for the Coordination of Humanitarian Affairs (OCHA) situation reports on the Haiti earthquake in early 2010. These documents are freely downloadable from the Web.

I also used, in anonymized form, parts of a practice spreadsheet that I had created for a workshop with LWF Nepal staff members.

I thank Cornelia Zuell for advice on applying a procedure to identify prominent terms. Thanks also go to Wouter van Atteveldt, Paul Buenau, Bruce Currey, Daniel Eriksson, Armando Geller, Hanspeter Kriesi, and Thomas Schwedersky for comments on an earlier version.

Introduction

Text analysis under time pressure

Recently I was asked to comment on a draft evaluation report. The report was the product that a team of two national and two expatriate consultants had created towards the end of their mission with a multi-sectoral rural development NGO. The national consultants had been hired to do two months' worth of preparatory work prior to the arrival of the expatriate members. The whole team spent between three and four weeks together in the host NGO.

The draft report went to considerable lengths discussing the NGO's current strategic plan and its affiliated sector lead documents. The size and complexity of the NGO required the team to look also at distinct sub-sectors within each major program. In tabular form, it contrasted the sub-sectors that the main strategy document enumerated with the sector plans regarding them that it could locate in subordinate documents. The effort, together with comments and interpretations, qualifies as text analysis. The coherence test of the main strategy design stretches across six pages, followed by about 35 pages of more empirical discussion of the major programs.

Towards the end of this section, the report takes a critical look at what the strategic plan mandated as cross-cutting issues for all programs – gender, governance, and environment. Two of the three issues were judged not seriously addressed in actual programming. The team criticized that the strategy was caught in "a web of different ambitions" and thus was "difficult to navigate".

In the conversations following the team's debriefing, "confusion about confusion" escalated. The team maintained its perception that the strategy was incoherent. The management did not understand the team's confusion about how the various elements worked together.

Regardless of the merits of these positions, as an external reader it struck me that the team's writing style did not give away particularly strong navigation skills either. There was no trace of any special text analysis tools that it might have used, during the preparatory period or later while together reading the strategy and plan documents. The progression of the report suggested that while the main strategic plan document had been meticulously parsed, subsequently the analysis of other plan documents had to proceed apace with processing notes from staff interviews and field visits. This had to do with the scope of the preparatory mission, which went directly into the collection of material that would speak to "achievements", without preliminary plan analysis. The full team, at some point of time, must simply have run out of time. It could no longer afford the luxury of text analysis with equally deep resolution in all the documents that it considered relevant. The field called, and staff wanted to talk.

That is a reader's impression; the report, of course, did not say it in those words. As many readers will know, time pressure is nothing unusual in program evaluation. It is not

unknown in slower-moving missions either, such as during reviews of monitoring systems or field research. In their book “Real World Evaluation", Bamberger at al. (2006) devote an entire chapter to coping with time constraints.

In the rest of this paper, I will talk of the types of people who perform humanitarian and development knowledge work under time pressure more or less interchangeably – the program staff in the first place, members of evaluation teams, users of monitoring data, field researchers, office-based economists, and other professionals. Common to many of their work situations is the challenge of having to analyze a number of documents, and to do so in short time.

The reasons why time is short are of interest to mission planners – the information arrives slowly, or the persons to work with it are expensive and are hired for as short a period as possible, etc. -, but are tangential here. The documents may comprise policy statements, project agreements, progress reports, interview notes, databases, images or other types. Time pressure may assail the extent and quality of the work that needs to be done with any of them. Generally, the analyst has several options to respond to the time pressure:

1. Selective attention to segments of the relevant texts and documents
2. Substitution of expert opinion for text analysis
3. Computer assistance

beyond normal search functions and word processing.

It may often be necessary to assimilate texts selectively and to point this out to principals and stakeholders, for example at key meetings during an evaluation. One hopes to catch up later or to agree on greater attention to this, and benign neglect of that, document. This is nothing unusual; experimental research in other contexts (e.g., Lurie 2004) has shown that information overload forces more selective acquisition.

Regardless of whether the analyst is upfront on his limitations or maintains a facade of paying equal and thorough attention to all relevant texts, detailed analysis may – in part or wholly – be replaced by opinion sought from persons familiar with the texts or their objects. This may be unavoidable or even desirable when the analyst himself cannot hope, from repeated readings and his own familiarity with the context, to resolve important inconsistencies that his first, incomplete reading revealed.

Yet, there remains the fundamental expectancy that the analyst take the texts of humanitarian and development organizations seriously, noting the importance with which principals and stakeholders commend them for consideration and analysis, and letting all of them, in principle, speak for themselves. Tools that promise to make working with text documents more efficient should thus reduce the need for selectivity and substitution. The personal computer, standard equipment of professionals, offers many such tools. This paper highlights some existing ones and offers a small addition of its own.

[Sidebar:] What is text analysis, and why should we do any?

Wikipedia, as of 5 June 2009, does not carry an article on text analysis. Instead, it enlightens us on "content analysis", often used synonymously, whose origins it locates in the 1930s.

Others point to beginnings much earlier in history. Rockwell (2003) finds text analysis' historic mover in Bible concordances. By the 13th century, concordances had become standard tools in the humanities. The first use of quantitative text analysis may be claimed by Swedes who, in the 18th century, took to counting religious symbols in songs (Popping 2000: 2). The advent of the computer gave concordance production a new thrust, albeit initially only for print concordances. Again, one of the first was in the realm of theology, the *Index Thomisticus*, started in the 1940s on index cards and not released on a CD until 1992.

Since then, interactive computer tools and the Web have enlarged our toolbox for dealing with texts from all realms and in many formats. One may thus consider text analysis as the totality of operations on texts that at some point are supported by computers, done with an analytic intention (as different from simply producing texts in their own right). But what is the threshold of effort and sophistication in order to qualify as text *analysis*? This is impossible to determine beforehand. For some, the definition thus remains elusive (e.g., Popping, op.cit.: 1) whereas others single out extraction of language patterns as the constitutive activity (Adolphs 2006).

Rockwell emphasizes that, centuries after the invention of the Bible concordance, some of the hermeneutic principles are still the same. Medieval scholars and contemporary text analysts both look for patterns of coherence in the text. They assume, for example, that the author uses a word in consistent meanings throughout its instances, and that the meanings can be clarified in the context.

Still, this does not answer *why* we should do text analysis. Goethe did not analyze texts. He read them. He wrote, and not badly. This raises a lot of questions, particularly about writing and creativity vs. analyzing texts and restating the obvious. Does text analysis really help to find the important insights that should come out of working with those who are the object, sometimes subject, of evaluations, research and other text-provoking encounters?

Moreover, it is questionable whether our cultural environment, driven by the Web, maintains coherent language. Search results are meta-texts that explode the unity of concepts that traditional human interaction presumed. Humanitarian or development-related text bodies are not exempt. Googling "refugee" throws up 16 million hits (as of 1st May 2009). "Refugee AND 'food relief'" returns over 80,000, a magnitude that does not surprise, given this essential need of refugees. "Refugee AND 'Cabernet Sauvignon'" stunningly returns over 4,000. Some of these are moralistic, contrasting the world of *savoir-vivre* to that of misery (for a blogger sample from the humanitarian world, Cortenraad 2000). Just as often, "refugee" parades a change in modern biographic identity, the *chef de cuisine* who fled the Rive Gauche for a wine paradise in California. Yet others are meta-documents such as library catalogues or collections of texts with disparate themes.

The example may seem abstruse, but the point is that these rogue meanings cannot be detected other than by actually reading the text. Yet, it is the very inability to read all relevant texts in a professional field that obliges us, like it or not, to use text analysis tools. They focus our ability to actually read a selection of texts that are relevant for the question at hand. One may thus speculate with Luhmann (1997: Chapter 2) that it is the excess of possibilities created by a new dissemination medium – the computer – that forces text analysis, as one of many devices to reduce complexity. In the humanitarian and development fields too, the proliferation of texts creates an attention economy in which the dilemma between attention (reading) and rejection is

mitigated, in small degree, by text analysis. It is a chore, not a free choice, for those made to scrutinize voluminous and diverse texts. But we can choose our tools.

Tools of the traveling expert

The laptop computer is one of the conspicuous trappings of the genus "*Sapiens barbaricus peregrinator*", or international expert. For most, these machines come equipped with an MS Office or open-source suite as well as with e-mail and Internet browsing applications. Generally, those are applications with which many of the local counterparts too work. For example, working together on data sets held in spreadsheets can provide a common vernacular right from the first days of collaboration between national monitors and expatriate consultants even when cultural and language differences in many other realms remain daunting.

Thus, at this level, file exchange and document collaboration may not pose many problems although different versions of the same application, memory limitations and deficient malware protection are frequent obstacles. Moreover, in locations with satisfactory access, the culture of working with Internet-based applications is fast catching up with that of richer nations, particularly within the Google and social networking families.

With much greater individual variability, computers carry applications from the realms of statistics, project management and logistics, GIS, image or Web page editing, bibliography, and local search. Work with any of these in a collaborative setting may remain limited to very few of the participants, either because the others do not have this software or are not skilled enough to use it for an effective team contribution. Even more rarely, participants may be able to collaborate using the same type of software for qualitative research, social network analysis, modeling and simulation, or for other exotic pursuits.

Job ads for longer-term in-country assignments usually specify expected computer and research skills. Occasionally, the terms of reference for short missions stipulate command of specific applications, for example of statistical packages for the analysis of survey data. For the much more common task of exploiting text documents for the manufacture of yet another text, both the experts and their principals seem to agree that normal Word processor and desktop search skills will do the job. The reason why specific text-analytical qualifications are rarely demanded is straightforward: Much more importantly, principals need to find personnel who 1. have the necessary language skills, and 2. can produce texts (notably reports) in the institutional formats required.

Text-analytical challenges arise when the texts to examine are voluminous or ill-structured in view of the time available and the purpose of the analysis. Yet outside of advanced text analysis applications, known and used by small communities of specialists (as in qualitative social research), there are few proven tools to yield rapid productivity gains. The more sophisticated ones demand a learning (and, for some, financial) investment that not all working in the humanitarian and development fields can make.

Purpose

The purpose of this paper is to add simple productivity tools for text analysis, by publicizing existing ones and by adding one that I created. "Simple" is a relative term. As the diagram in the Summary section suggests, the suitability of the tools depends on the skills and equipment level of the intending user. Also, I assume a kind of working environment that developing country organizations will not everywhere offer for computer-supported text analysis: that the analyst actually can acquire the documents digitally.

The need for these tools arises in two distinct situations:

- The analyst is comfortable with classic interpretive (i.e., non-statistical) approaches, focused on the intent of the document authors, the texts' internal logic, and the relevant professional and audience contexts. The need is for rapid navigation and comparison inside and between texts. For a typical situation, think of the briefing papers that an evaluation team receives, some together with the terms of reference, some at the project site.

- Second, purely interpretive approaches break down, because of the volume or structure of the text material, or are not chosen in the first place. The analyst, with the help of statistical tools, investigates distributions and correlations of text elements that carry meaning, notably words and terms. Use of such results may range from navigation to exploration to testing of hypothesis. For example, a federated international NGO may produce major policy documents centrally. Over time, local policy adaptation and implementation are reported in numerous documents created in the participating member organizations. Their volume defies direct interpretive access.

The tools presented here serve both the strictly interpretive analyst and the statistically minded, in differing degrees and mixtures. In fact, they should help building bridges between methodological communities. The reader may choose which of the three tools, if any, will possibly benefit his work. These three are located at different heights of a learning curve. The easier ones work equally well regardless of whether the more demanding ones are adopted or not. The more demanding ones demand skills in spreadsheet and statistical applications.

Organization of the paper

The main part of the paper presents these tools:

- TextSTAT is a concordance, word frequency and document navigation tool, created by an external researcher. It appeals to the interpretive researcher who finds navigating through hardcopies or in word processors slow.

- The popular spreadsheet application MS Excel can be put to uses that straddle both traditions, particularly when the number of distinct texts is significant. I

wrote a composite function to flag terms in text stored *inside* the worksheet as well as a macro that extracts term frequencies from a set of *external* Word or .txt documents. I also offer a macro that computes a *term association* matrix, which can be used to graph out relationships among concepts. Moreover, Excel 2007 users find a convenient, easy-to-use tool in NodeXL to present text structures that can be visualized in network graphs.

- Wordscores is a collection of routines written for the statistical application STATA. It creates word frequency tables extremely fast, and with the option of an important text preprocessing operation (stemming) that facilitates subsequent statistical analysis.

I describe for each tool the situation in which it makes, hopefully, text analysis more efficient (and, in fact, more effective by leading to discoveries that would not likely be made otherwise), what it does, and, to a point, how it does it. The more arcane technicalities of Excel macros and Wordscores are banished to appendices, but the reader wishing to adopt the tools should read these. Sidebars in the main body add definitional clarity and offer examples of statistical analyses.

In a further section, I briefly discuss tools that are yet higher up on the skills gamut, and which, admittedly, are beyond my competence: I do know about, but do not currently use, advanced programs for text analytics (analytics, not analysis!) or for qualitative research. The learning curve for these is more arduous than for the quick-to-learn tools offered here.

What is this paper not about? If it focuses on tools, it is lean on theory. My choices are colored also by the want of linguistic education that limits explorations of more demanding text processing and data mining applications. For example, I do not develop my own rationale for working with word frequency tables; I simply report someone else's reasons why we may want them. My excuse is that the reader will have her own rationales, but may look for better tools to carry them out. Readers anxious to see substantive examples in the humanitarian and development field may want to look up Goldman's research into the "greening" of the World Bank (Goldman 2001) or my discussion of the dizzying career that the empowerment concept achieved (Benini 2008: 22-32).

[Sidebar:] Some useful definitions – Document, words, terms

- A **document** is a sequence of sentences
- A **sentence** is a sequence of tokens
- A **token** is the smallest unit of text
    - E.g. words, numbers
- A **term** is a token or a set of tokens with a semantic meaning
    - E.g. single token, proper names
    - "San" is a token, but not a term
    - "San Francisco" are two tokens but one term
    - Dictionaries usually contain terms, not tokens
- A **word** is either a token or a term – depends on the context

- A **concept** is the mental representation of a real-world thing
    - Concepts are represented by terms/words
    - Terms may represent multiple concepts: homonyms
    - Concepts may be represented by multiple terms: synonyms

(Leser 2008: 32). Outside of text analysis, it is common to use the word "document" – sorry, the *term* "document"! – also to refer to information bodies without sentences, for example, spreadsheets or photo albums, and to the computer file in which the document is opened or saved.

Tool #1: TextSTAT

Among a score of free text analysis programs[2], TextSTAT (Hüning 2010) stands out by its simplicity, clarity, ability to read several file formats, search and link ability, as well as the ability to produce and export word frequency lists. TextSTAT is ideal for the evaluation team member working through multiple text documents that she needs to analyze in conceptual areas for which salient key words exist. These may already be known or may be detectable from their frequent use in these texts. TextSTAT's features and functions can be learned in less than half an hour.

Currently, TextSTAT handles ASCII/ANSI texts (in different encodings), HTML, MS Word and OpenOffice files. It requires the user to create a new (or open an existing) corpus file. Formally the corpus is simply a holding vessel for all the text files that we wish to view together, saved with the file extension ".crp". Substantively it is the collectivity of all those texts, with their words, concepts and other linguistic relations contained in them.

[Sidebar:] Text corpus

Linguists will be horrified at my pragmatic definition that the corpus is "simply a holding vessel for text files". I am highlighting this term only because the TextSTAT user will encounter it.

For those wanting a more satisfactory explanation, here is a segment from the Wikipedia entry on "text corpus":

> *In linguistics, a **corpus** (plural corpora) or **text corpus** is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe.*
>
> *A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora.*
>
> *In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags. Another*

[2] Brief descriptions, and links to the sites, of other programs can be found at (DiRT 2009b) and, for freeware only, (Altman 2008). For a similar listing of qualitative data analysis programs, see (DiRT 2009a).

example is indicating the lemma (base) form of each word. When the language of the corpus is not a working language of the researchers who use it, interlinear glossing is used to make the annotation bilingual.

Some corpora have further structured levels of analysis applied. In particular, a number of smaller corpora may be fully parsed. Such corpora are usually called Treebanks or Parsed Corpora. The difficulty of ensuring that the entire corpus is completely and consistently annotated means that these corpora are usually smaller, containing around 1 to 3 million words. Other levels of linguistic structured analysis are possible, including annotations for morphology, semantics and pragmatics.

Corpora are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics, speech recognition and machine translation, where they are often used to create hidden Markov models for POS-tagging and other purposes. Corpora and frequency lists derived from them are useful for language teaching.

Although I used a (in the event, British) corpus for one of the statistical examples (see page 38), dealing with such large external text bodies is not at all necessary while using TextSTAT and the other two tools proposed in this paper. The reader may want to know, however, that there are large-corpus derived word frequency tables accessible. The American National Corpus[3], for example, has frequency data for both written and spoken American. The table on the written version was extracted from an 18.3 million word corpus and carries over 290,000 word form entries.

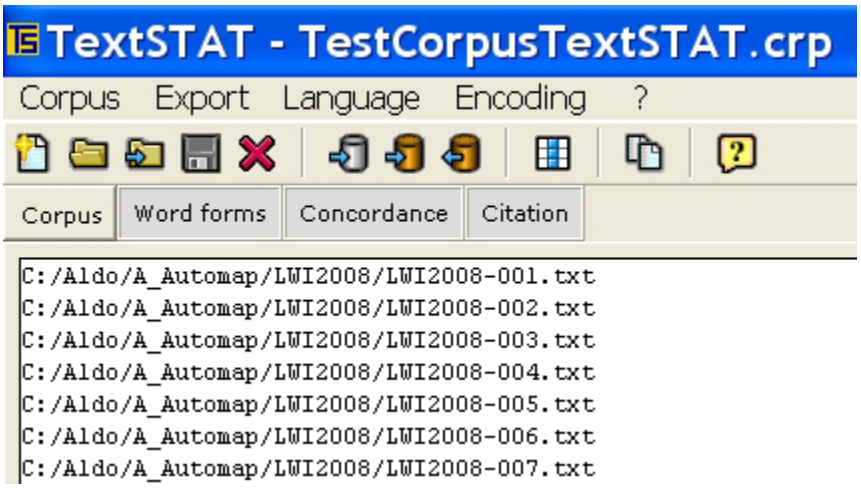
TextSTAT supports four operations, each marked with a button that opens a different view:

- Corpus
- Word forms
- Concordance
- Citation

The corpus appears as a list of the files that we choose to include, the file names complete with their directory paths or URLs.

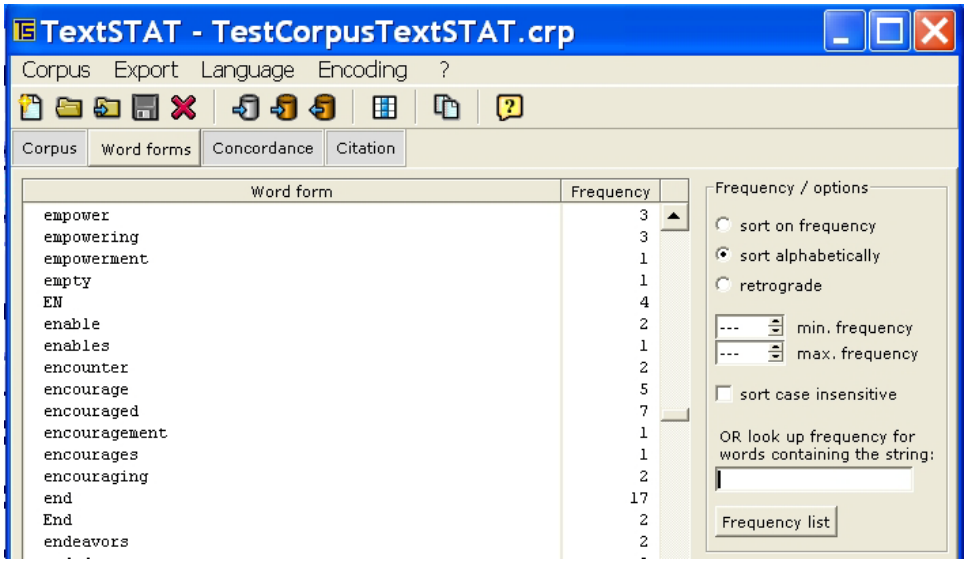
[3] http://americannationalcorpus.org/index.html

Figure 2: The Corpus view in TextSTAT



"Word forms" offer a choice between a listing of all word forms that occur in the corpus or one restricted to words containing a particular string. For each form, the number of occurrences (frequency) is calculated. The frequency list can be exported as a .csv file and can then be treated in a spreadsheet or statistical program[4]. In the message box at the bottom, statistics are given of the number of word forms/types as well as of words/tokens in the entire corpus.

Figure 3: The "Word forms" view in TextSTAT

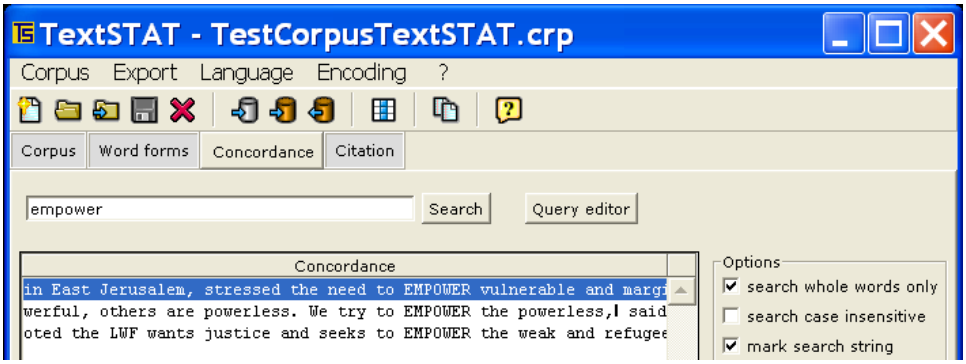


The concordance – the listing of occurrences of a word within their immediate contexts – can be established in two ways, either by double-clicking the word in point inside the Word forms view, or by using the search box in the Concordance view itself. The former option makes the tool particularly attractive; it allows repeatedly moving between Word

[4] The export option "Frequency list > MS Excel" (direct) worded not work for me.

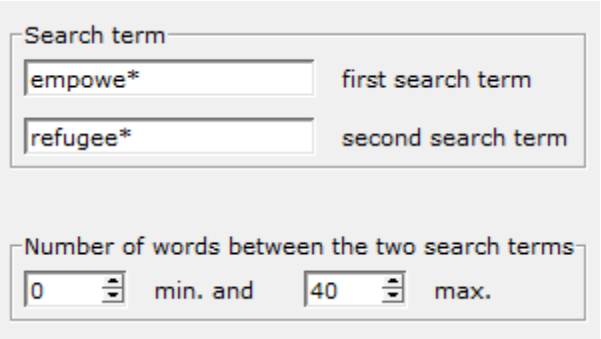
forms and Concordance, following a context element for its own occurrences in other context. Also, the concordance can be exported to a .txt file or directly opened in Word.

Figure 4: The Concordance view in TextSTAT



In the Concordance view, searches can be refined with the help of the Query Editor. Two search terms can be combined, with the minimum and maximum number of words between the two terms specified. Wildcards (* and ?) are allowed.

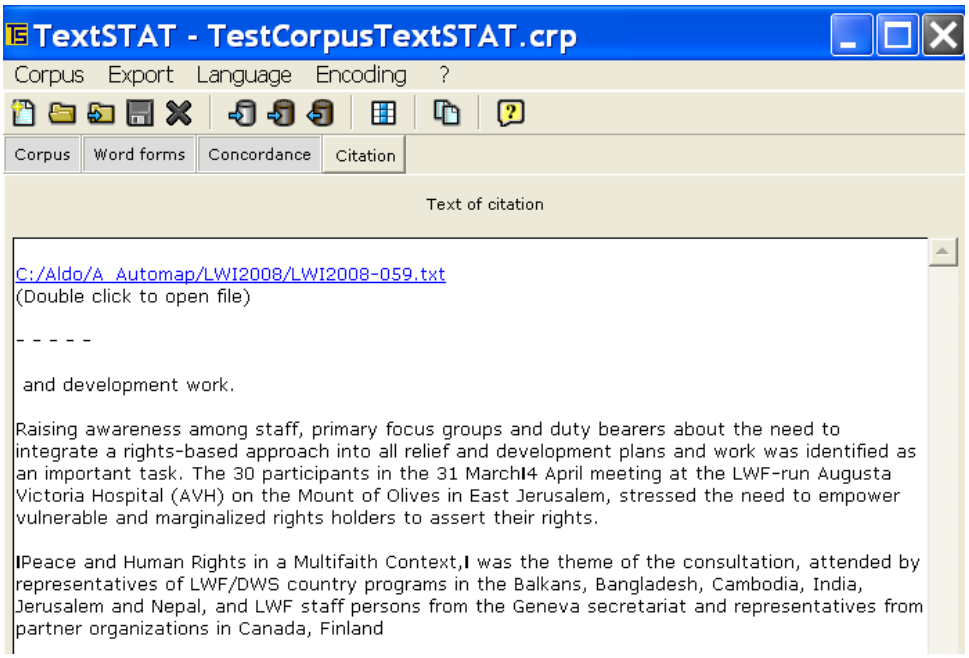
Figure 5: The Query Editor in TextSTAT



Double-clicking a concordance element will open it in the Citation view. This has two elements. In the lower half the short concordance text appears highlighted in red[5] and surrounded some more of the context to the left and right (in black font). On top, we find the hyperlink to the text file of which the citation is part.

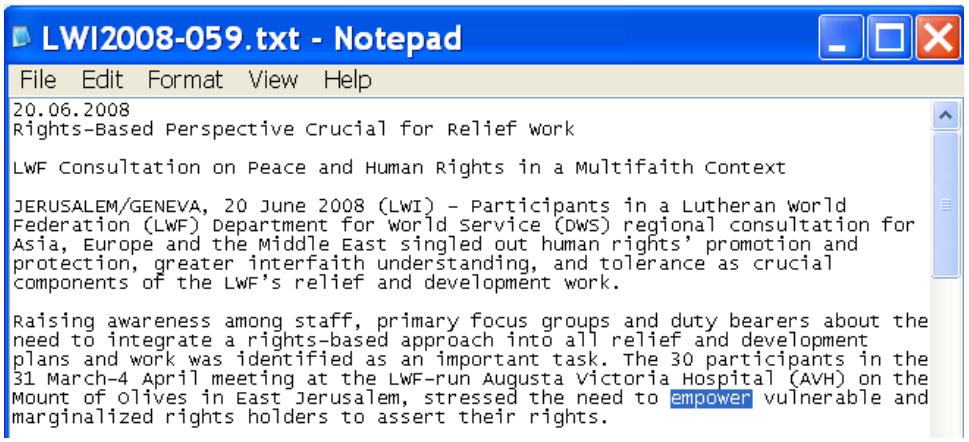
[5] The red does not show in the screenshot image below, but it does in the application.

Figure 6: The Citation view in TextSTAT



In this instance, a simple text file is opened.

Figure 7: A corpus document opened from TextSTAT



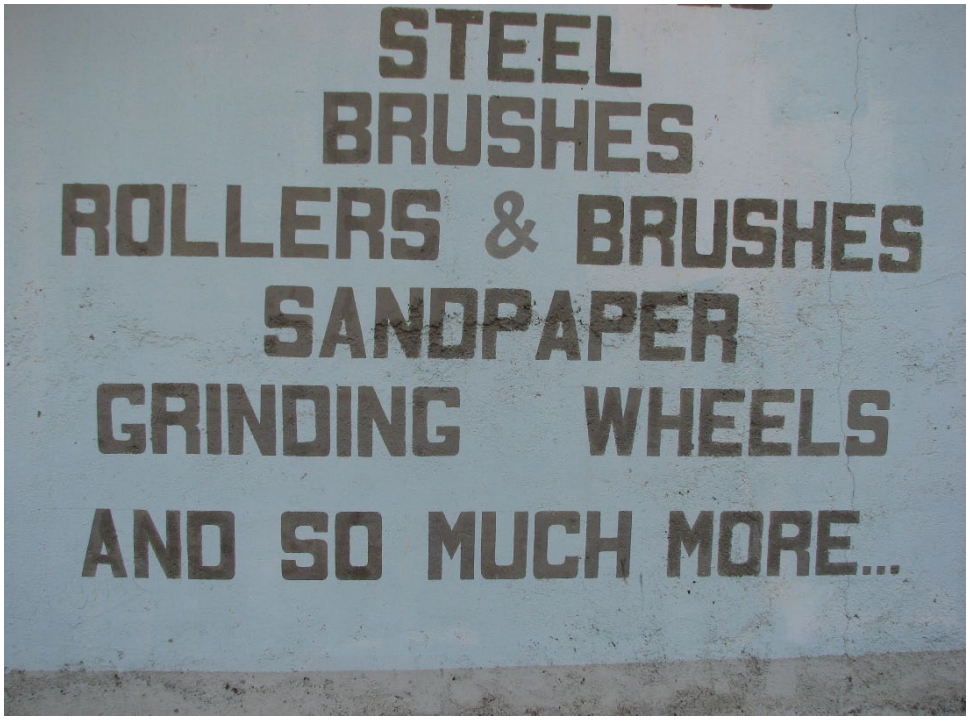
Note: "empower" highlighted by search in Notepad, not automatically.

Two caveats should be voiced. First, it is not advisable to open files that are members of an active TextSTAT corpus in their own applications, and even less so to modify and save them. Doing so, i.e. opening a .doc file in Word, making changes and saving them while it is being used in TextSTAT, may send TextSTAT crashing. To avoid those problems, while at the same time working on some of the active corpus files, one should make working copies; images in the copies to be loaded should be deleted.

Second, TextSTAT does not read .pdf files, thus excluding frequent types of documents (e.g., NGO annual reports) that one may find in this format only[6]. In this situation, two choices may be considered: 1. conversion into text or Word format (with the risk that text becomes garbled, particularly if it is in column format) and inclusion in the TextSTAT corpus, or 2. placing the needed .pfd documents in one subdirectory and using the "Full Reader Search" option in the drop-down menu of the "Find" box in Adobe Reader, outside of TextSTAT.

That said, TextSTAT is still a highly efficient tool when we compare its useful multiple functions to its cost (it is free), extremely short learning curve, low demand of computer resources, and possibility to use it in a team and/or with local counterparts[7].

[Sidebar:] Steel brushes and much more



This wall advertisement, three meters tall, stares at passers-by in Dili, East Timor, a short distance from the title page scene. Though drab and unappealing to all but to the cognoscenti of the building trade, it confirms two points for the text analyst:

---

[6] The most popular PDF reader, Adobe Reader, has a facility (File → Save as Text) to convert unprotected .pdf documents to .txt ones. The conversion is slow.

[7] Once the set of documents to analyze has been determined, the team member anxious to let others participate may simply copy them to the subdirectory "Files" within a directory "OurTextSTAT" on the c: drive. He adds them to a corpus in his TextSTAT and saves the corpus in the subdirectory "OurCorpus" as, e.g. "EvalDocCorpus.crp". In a third sub-directory "Installation", he places a copy of the TextSTAT installation file (currently "TextSTAT-2.exe", 3.2 MB). The whole project can then easily be shared among the intended participants, who may keep TextSTAT on their machines for their own future projects.

- **Text and social structure are correlated.** A small minority in Dili reads English. As far as I could see, there were no Tetun, Bahasa Indonesia or Portuguese versions on display. The poster appeals to the clientele with the greatest purchasing and - given their positions in foreign security forces - oil companies and UN reconstruction programs, decision making powers to buy these goods. In this sense, the advertisement itself is the result of a rational text analysis – of such texts as inventory lists and customer orders.

- The *"and so much more"* is important. Inviting potential buyers to come to the store and find out for themselves all the treasures and surprises of the building supply trade, the expression signals **surplus meanings.** The builder who needs brushes needs paint too. We have it, or at least we can order it for you. The wealth of meanings emanating from the unwritten context (or from context elsewhere in the corpus) is one of the strongest arguments for interpretive text analysis and against statistically supported analysis.

Or at least so it was believed for a long time. Modern computing power can, to an extent, overcome this limitation by extrapolating from the relations inherent of large text corpora. Development and humanitarian workers do this whenever they use Internet search machines. But the algorithms are opaque to all but specialists, and value and effort can be out of balance for these kinds of supporting searches. One of the consequences for text analysis is that the gulf between interpretive and statistical workers may become less deep, but it will not be filled completely.

Tool #2: Text to numbers in Excel

The frequency tables that TextSTAT produces may stimulate a quantitative treatment of the concepts in a corpus. A question that may arise in various perspectives is how the relative frequencies of concepts vary across the member documents of a corpus. If, for example, the documents represent a sequence of reports from the same source over a number of reporting periods, one may ask how the prominence of certain terms has changed over time, or in response to important events that took place at some known point in time. Similarly for documents produced for one period, but by various units of an organization, such as the country programs of a large NGO federation. Admittedly, word frequencies are only part of the overall wealth contained in documents (or speech, for that matter), but they, plus the associations among the concepts used, are important dimensions of the total information or knowledge that the documents transport. Many interesting qualitative questions may arise from noteworthy patterns detected in these statistical relationships.

Word frequency tables can be exported to spreadsheets, and spreadsheets are highly versatile in the kinds of variables that they can handle (including text elements up to a certain length). It thus makes sense to look to this kind of application for the next higher level, in abstraction, of text analysis. Here, I investigate some selective techniques using the spreadsheet application MS Excel. Excel is widely used in the statistical and NGO monitoring communities. That, together with its inbuilt text functions, makes it a first-order candidate for the treatment of such problems. At the same time, one has to realize that Excel is not conceived of primarily as a text analysis program.

Two basic situations need to be distinguished:

1. The text material of choice may reside in the spreadsheet already, and will thus be directly accessible to Excel's own tools.
2. Alternatively, the texts may reside in external files, typically in different formats such as .txt (text) or .doc (Word files), in which case Excel needs an extraction tool to access these documents and return the information of interest in a useful format.

In both situations, the product is a term – unit (external document or internal cell content) table, either returning frequencies (how many times the term appears in the document) or binary occurrences (whether it appears at all). While much can be learned from inspecting such tables, depending on the question at hand the researcher may want to compare term-to-term relations, notably co-occurrence. The instruments presented here address both term-unit and term-term relations.

Spreadsheets with text variables

Excel text functions are most often used for data manipulation in tables that fill database functions. For example, if names were imported into a field that holds both first and surnames ("Jack Miller"), one may want to distribute these elements to separate fields ("Jack", "Miller"). Most of these database problems are tangential to the concerns of text analysis; they operate, so to say, at a lower level of data preparation. In a not-too-rare exception to this rule, an Excel text field may hold the uncoded response to an open survey question. These entries can be manually coded into a separate field, by persons looking at each entry and then determining which category fits the open-form response best. But one could also imagine that a sentence like *"In our village, HIV has caused most young people to migrate to the cities"* should be noted for several elements, the salience of *HIV* as well as that of *migration* (and maybe *youth* as well). Extraction to more than one field may thus be desirable. It can be achieved using suitable Excel text functions.

Excel cell content can be wrapped so that the entire text is visible in the given cell, spread over several lines. This feature makes it attractive to use Excel rather than Word to tabulate mixed numeric and text information when calculations are needed. Two examples may suffice:

In program monitoring, fields in such tables can be used to hold short text descriptions of events that are of a repetitive nature, yet diverse enough to resist pre-coded entries. Other fields may be numeric, holding information on dates, number of participants, costs or other financial quantities, or standardized text such as the names of administrative and organizational entities. Logbook-like running tables are suitable particularly for activity reporting in federated structures, such as when an international NGO supports a number of regional CBOs, which in turn work with a host of local grassroots associations. The latter may generate a large number of events within a relatively small number of event types (seminars, protest marches, individual assistance grants, etc.) and the attendant numerous elemental entries in their reporting. At the same time, in the upstream reporting, one wishes to retain the local characterizations of the events and to avoid aggregating out the special flavors and unique elements too early. Similarly to open-ended survey

questions, “unstructured” spaces in the reporting forms (“issues tackled”, “impact on members”) may capture some of the local diversity. Such arrangements often degenerate into trivial routines (“strong impact”, “ditto”, “ditto”), but with appropriate training and careful feedback, some nuanced event-based information can be kept flowing. The challenge then is to appropriately analyze it.

Second, some NGOs implement amazingly large numbers of projects that they subcontract from donor agencies and national governments. The project names used in funding agreements are often quite verbose, chaining together a number of regional, sectoral, target group and latest development-philosophy terms. Long tables detailing donor, start and end date, grant size, and name of project may conceal clusters of socially and substantively similar projects. Dominant project types can be detected if the terminologies used in project titles are suitable parsed. As an observer of NGO programs in Bangladesh once mused, “asset transfer” might likely be written into the same project title together with “poverty *reduction*” if the grant was smaller than one million dollar, but might more likely go hand in hand with “poverty *elimination*” claims in substantially larger programs. If one cares, Excel functions can flag these differences in the text elements of project listings.

The analyst may use any number of strategies to cope with this kind of data within Excel; there are few a-priori limits to analytical ingenuity. Note, however, that cell entries in Excel 2003 are limited to 1,024 *displayed* characters, and in version 2007 to 32,767. More constricting, particularly for term matrices, is the maximum of columns in a sheet, which for Excel 2003 is 256. Depending on the user’s installed version, text storage and display strategies may thus differ.

This difference is important for the possiblities that Excel offers for text analysis. We will deal with two example applications that are feasible in Excel 2003. The first concerns log-book like tables in which the text elements to be analyzed reside in the spreadsheet cells. The second extracts the presence of terms from a multitude of external text documents, resulting in word count or co-occurrence tables in Excel. Subsequently, I illustrate the more generous potential of Excel 2007, including the visualization of certain results in network graphs.

Analysis of logbook-like tables

We assume here that we deal with text stored in cells of Excel worksheets, version 2003 and earlier. Long text segments, if held in one cell, would make logbook-like table presentation unwieldy.

Text elements composed of fewer than 1,024 characters (including spaces) and cleaned of special characters such as paragraph marks can be displayed in one cell of an Excel 2003 spreadsheet. As an ambidextrous tool, Excel naturally is indifferent about the purposes and meanings of the information it can hold. For the user, however, the space limitation curtails the variety of texts that can be efficiently handled within this format. Excel is

attractive to the extent that short free-text elements need to be stored and analyzed side by side with numeric and standardized-text elements. Besides the flow of events calling for logbook-like reporting, other information needs may create this type of situation[8]. In relatively small and simple surveys, data including the response to open questions (as long as interviewees do not elaborate effusively) fit the Excel bill.

Excel offers more. Of particular interest to the hurried researcher, it facilitates the movement from text to statistics, and hence back to (selected subsets of records with) text.

A contrived example may illustrate the mechanics involved. In a federated network in Nepal, a monitoring workshop produced a fictitious event-based reporting table, of which a small segment (the first two records) is shown here[9]. Besides a small number of standardized fields (the names of the intermediary NGOs [termed "citizen based organizations", CBOs] conducting the events, event dates and durations, the number of participants), much of the essential information was kept in the text field "Comment". The central NGO was anxious to evaluate this report in a flexible way that would produce some statistics of interest, but also keep the information on individual events readily accessible.

Table 1: A log-book like activity table

Rec. No.	CBO	Start	End	Duration (Days)	Particip.	Comment
1	CBO5	2-Jan-07	13-Jan-07	12	16	Organizational experience translated into improved practice. For this reason, an accompanying training component is foreseen.
2	CBO1	4-Jan-07	18-Jan-07	15	4	Self-help groups and networks: The empowerment programme will enable Dalit participants (60% women) to improve their status and living standards through social literacy-empowerment.

This was done in three steps:

1. Flagging, in separate fields, the presence of terms of interest in the comment field
2. Creating Pivot tables of key statistics, by organizing NGO, for any combination of selected terms
3. Tables filtered to only those records meeting the interesting combination of key terms and Pivot table categories

[8] For most survey researchers, Excel may not be the application of choice, but in developing nation NGOs, survey data is often entered into spreadsheets, if only for lack of support for other applications.

[9] Local associations were anonymized, and empirical report elements replaced with simulated figures and, for text, with snippets from a planning document of the coordinating NGO.

Step 1 relies on the most involved mechanics. This is dissected in the appendix.

In step 2, we use the Excel workhorse for tabulations, the Pivot table facility, to customize statistics to the presence or not of a search term or a combination thereof. In the Nepal exercise, 10 citizen-based organizations (CBOs) reported 57 events. 32 of these, from 9 CBOs, involved Dalit issues (groups of people in South Asia considered untouchable). This result was speedily obtained by placing the "Dalit" binary variable in the page field of the Pivot table and setting its value to "1". The table shows the total number of events of participants and of participant days under each CBO. Note that CBO No. 6 did not conduct any Dalit-related event and does not figure in this table.

Table 2: Summarizing log-book information in a Pivot Table

Training	(All)
Dalit	1

Reporting CBO	Events	Participants	Participant days
CBO1	7	42	1,154
CBO2	1	3	141
CBO3	5	72	2,806
CBO4	2	7	127
CBO5	5	71	765
CBO7	4	48	1,279
CBO8	4	23	232
CBO9	2	23	523
CBO10	2	7	67
Total	**32**	**296**	**7,094**

The third step takes us back to the full information for specific subsets of records, including the text fields. Assume that we have a special interest in the Dalit-related events that CBO No. 4 conducted. As the above table shows, there were two such events in the reporting period. By double-clicking a results cell (e.g., the cell with the 2 events), Excel creates a new table holding all the information for the subset with the desired combination of properties: 1. Training as well as non-training event, 2. Dalit-related events, 3. those reported by CBO No. 4. The result shows, among other things, that neither of the events was a training event, as far as the comments tell us:

Table 3: A query of the log book called from the Pivot table

Event No.	CBO	Particip.	Comment	Training	Dalit
19	CBO4	3	This CBO is a recognized supporter of poor, marginalized and disadvantaged communities (Dalits, bonded labour, indigenous groups); known for its significant geographic outreach.	0	1
41	CBO4	4	This event was part of Dalit Empowerment Programme for Western and Central Nepal.	0	1

The reader may find this two-event example trivial and unnecessary, but Excel Pivot tables can instantly produce sub-tables with hundreds or thousands of records, including their text fields.

This three-step procedure in Excel has a remote, incomplete parallel with the progression of windows in TextSTAT: the table of event records takes the place of the corpus with its constituent documents; the Pivot table resembles the Word list with associated frequencies; and the sub-setted specific tables of the third step function like the concordance. The Excel workbook is both more versatile and more limited than the TextSTAT-managed corpus: it allows us to correlate variables from outside the text fields (categorical, numeric, dates) with text represented by markers for terms of interest. But these terms have to be pre-defined by the researcher and are limited in the numbers that are practical for analysis. Exploring a variety of Pivot tables and inspecting sub-setted tables of interest will, as a rule, be slower than the swift navigation among corpus documents and key-words-in-context that TextSTAT permits.

Extracting from external text documents

TextSTAT creates, as we have seen, word frequency tables for the entire loaded corpus. However, for comparison purposes, it may be necessary to extract the word frequencies for each of the documents in the corpus and to present this information in a format that supports visual inspection as well as statistical analysis. Some readers may wonder why they should bother to look at word frequencies at all unless these tables are directly linked to some other function such as the powerful concordance feature in TextSTAT. The number of occurrences in which a word (or other searchable expression) appears in different documents is de-contextualized information. As such, it is not obvious how it can possibly aid the interpretation and comparison of texts in policy, programmatic and other relevant contexts.

The motivation for studying word frequencies across documents is succinctly given by Duriau et al. (2007: 6):

> *At its most basic, word frequency has been considered to be an indicator of cognitive centrality or importance. Scholars also have assumed that the change in the use of words reflects at least a change in attention, if not in cognitive schema. In addition, content analysis assumes that groups of words reveal underlying themes, and that, for instance, co-occurrences of keywords can be interpreted as reflecting association between the underlying concepts.*

From here, one can easily anticipate some of the major types of word frequency analysis:

- Documents within a corpus – such as the half a dozen or so basic policy documents and key reports that an evaluation team may receive in its initial briefing meetings – can be compared among themselves, along such distinctions as plans vs. reports, or earlier vs. more recent reports.

- The word frequencies of the corpus at hand may be compared to (usually much larger) external corpora that represent language use in the wider society, or in a different organizational or professional community. I mentioned the American National Corpus. One of the most popular is a set of over 20,000 news stories known as the *"Reuters-21578 text categorization test collection"* (Lewis 1997) and its successors.

- Third, the co-occurrence structure of significant terms in the corpus may be of interest. The co-occurrence may be expressed in manifest ways, such as through networks of terms that appear together in many documents. For example, within a rights-based approach to development planning, "duty bearer" is likely to occur in the vicinity of "rights holder". Co-occurring terms may also be seen as expressing latent concepts for which none of the corpus terms covers fully, and which need to be interpreted by the researcher relying both on theory and on the known context. Such a challenge might arise in long-term perspectives on programs that have evolved along the relief-to-development continuum.

In many analyses that fall under those approaches, the word frequency table, or as it is sometimes called, "term-document matrix" (Feinerer, Hornik et al. 2008: 10), is a critical semi-product. We exemplify this with a small segment modified from the frequency table for the set of 76 Web articles that the information service of the Lutheran World Federation (LWF)[10] published in the first half-year 2008. In the remainder of the paper, I call this collection the "LWI corpus".

Table 4: A segment of the term frequency table for the LWI corpus

Terms	LWI2008-007.txt	LWI2008-034.txt	LWI2008-059.txt	*AllOccur76Docs*
Lutheran	72	2	3	*415*
Lutheran World Federation	1	1	1	*78*
empower	0	0	3	*7*
human	0	0	8	*123*
human right	0	0	7	*19*
camp	0	1	1	*21*
refugee camp	0	0	1	*3*
TotalWordsInDoc	*2,271*	*316*	*1,112*	*53,191*

The corpus of 76 articles is made up of 53,191 separate words and other string expressions such as numbers. The table gives the occurrence counts for a small number of terms that are salient in this federation of member churches and, at the same time, international humanitarian and development agency. The three documents in the above table were chosen for illustrative purposes only:

[10] http://www.lutheranworld.org/News/Welcome.EN.html.

Table 5: Titles of three sample documents in the LWI corpus

No.	Document title	Date
-007	Global Increase in LWF Churches' Membership Pushes Total to Over 68.3 Million	15 Feb 2008
-034	LWF General Secretary Pays Tribute to Theologian Lukas Vischer	31 March 2008
-059	Rights-Based Perspective Crucial for Relief Work. LWF Consultation on Peace and Human Rights in a Multifaith Context	20 June 2008

The differences in the frequencies of the selected terms are conspicuous, but are easily explained in the light of the article themes:

- The first two deal with member church and Lutheran communion-related matters. The first, concerned with church growth, uses the term "Lutheran" no fewer than 72 times. By contrast, we find no instance whatsoever of the terms that we may associate with the "humanitarian persona" of the Federation, such as "empower", "human rights", "refugee camps".

- In the second appears an occurrence of the term "camp", but close inspection reveals that it has nothing to do with refugee camps or political camps. Rather, it is a 4-letter string contained in the word "campaign", as in "the churches' campaign on climate change".

- The third, focusing on rights and relief, uses the word "human" eight times, seven of which in the expression "human rights" (and one as "human resource manager"). But as the rightmost column shows, the almost exclusive use as "human rights" is not typical of the corpus as a whole: of the 123 instances of "human" in the 76 articles, only 19 complement to "human right".

What does this tell us of general interest? First, as one could gather from a chance reading of a few articles totally unassisted by any statistics, the intra-corpus variability in conceptual emphases is high. At the same time, some frequent markers lace most documents (in this event, "Lutheran" above all others). This reflects the normal tension between redundancy and diversity, coherence and novelty that keeps the readers motivated.

Second, the naive use of frequency tables composed solely of one-string search terms may be misleading. Technically, purely lexical ambiguity as seen between "camp" and "campaign" can be controlled. One can, for example, program the frequency table algorithm to count whole words only. More importantly, and taking greater effort to ensure meaningful text analysis, composite concepts such as "refugee camp" and "human rights" may need to be deliberately included in the term list. Similarly, so-called named entities such as the "Lutheran World Federation" need to be recognized, together with their acronyms ("LWF"). Depending on the question of interest, one may want to calculate occurrences for "human right" and "human, NOT: human right" separately.

With these cautionary remarks in mind, Excel users wanting to produce their own word frequency tables for a set of search terms and a set of documents may find the macro in the appendix helpful. *"FreqGivenTermsMSWorddocs"* calculates the occurrences of each search term for each of a number of Word documents saved in the same subdirectory. Terms can be multi-word. At the bottom of the frequency table, the macro records the total number of words (in the definition of MS Word[11]) for each document. This statistic is needed when relative frequencies (such as occurrences per 10,000 words) are to be calculated.

The user has to set two parameters in the macro, which he accesses in the VBA editor by pressing Alt+F11. These are: the path to the subdirectory that holds the documents, and the number of the bottom row of the list of search terms. The search terms must be held in column 1 of the worksheet that receives the frequency table.

The macro extracts from Word documents only. I included a companion macro for .txt documents. The reader will not find it hard to modify it for collections of .htm documents.

The name of the macro "Frequency of given terms in MS Word documents" reminds the user that he must provide a list of search terms beforehand. Of course, one can also create the complete word list of the entire corpus using TextSTAT and export it to Excel. But the execution of searches for perhaps several thousand word forms and of the transfer of all counts to the spreadsheet may be prohibitively time-consuming. It is better suited for selective approaches. The beauty is that it handles multi-word terms and named entities as well. And the benefits of team collaboration among members whose computers are equipped with Excel apply equally to the area of application of this macro. A member who has never before worked with macros can, in less than five minutes, have this one installed and explained by one of his more Excel-savvy colleagues.

[Sidebar:] Automatic term extraction

TextSTAT's frequency list returns only one-word terms. This is an unfortunate limitation; for single words do not render all of the most important terms in a text, and certainly not in humanitarian and development-related ones, given the complex ideas that their language tries to convey. Thus, in TextSTAT, the word list would reflect both statements: "she lost the use of her right hand" and "a grave violation of this human right" by adding one instance each to "human" and to "right". In order to produce counts of "human right" or "right hand", one would have to work in its query builder, one term at a time.

A tool developed by computational linguists and called "term extraction" can provide helpful supplementary information in the shape of multi-word terms that a statistical algorithm identifies in text documents. A small number of term extractors are available on the Web; the user uploads a text (or corpus of texts) and in turn receives a list of the most important terms and their values on one or several linguistic metrics.

[11] The differences in Word count between MS Word and TextSTAT are stupendous and, so far, not explicable. MS Word counts 53,191 words in the LWF corpus of 76 articles. TextSTAT reports 44,577 words or tokens from 6,232 word forms or types in the same corpus.

TermExtractor, an algorithm hosted by the University of Roma "La Sapienza" (Sclano and Velardi 2007; LCL 2009), returned a list of two and three-word terms that it ranked as particularly relevant in the Oxfam International 2007 annual report (Oxfam 2008). The following table renders terms that earned the highest weights. The weights are a combination of four metrics: domain relevance, domain consensus, term cohesion, artificial frequency.

Table 6: Most highly weighted terms extracted from the Oxfam annual report

Term	Term Weight
partner organization	0.86
strategic plan	0.86
joint work	0.85
developing country	0.79
info oxfam	0.78
climate change	0.77
essential service	0.76
gender justice	0.75
basic social service	0.69
life-saving assistance	0.68
social service	0.68
sanitation facility	0.66
donor government	0.66
civil society	0.66
developing world	0.66
honorary president	0.65
water supply	0.65

Practically, such term lists can be useful in narrowing the perspective of an ongoing text analysis, both substantively and formally. Did Oxfam's strategic plan emerge from joint work with its partner organizations, as an immediate reading of the three top returns suggests? Formally, if we were to investigate a corpus of several texts (not necessarily all from Oxfam), we could use it in the production of a term frequency matrix of the kind that the next section enables.

Term-to-term relations

Once significant terms for the investigation of a text corpus have been identified, it may be enlightening to find out which tend to occur together and how often. Obviously there can be different measures of interest. One may ask, for example, in what fraction of the documents term A as well as term B occur.

However, texts transport meaning through a sequence of terms. The terms appear in networks of associations that make it likelier that, say, A is followed by B rather than by C. These associations have a direction, making "A followed by B" likelier than "B followed by A". Thus, if "followed by" means "occurring later in the same phrase", "human rights" will be met more frequently than, say, "the rights of human beings".

The instrument offered here uses a different interpretation of co-occurrence. Rather than being concerned with the order of terms within one text unit, it investigates: *"In documents in which term A occur, what is the chance for term B to occur?"* This question too admits different numeric responses. The AssocMatrix macro uses an existing term-

document table to calculate, for each pair of terms A and B, the fraction of corpus documents with A occurring in which also B occurs. Technically, this is the conditional probability p(B|A) for simple co-occurrence.

In his "*Semantic Network Analysis*", van Atteveldt (2008: 71-73) proposes a more conservative measure that takes into account multiple occurrences of A and B in the documents[12]. If the corpus under study is regarded as a sample only of a larger population of relevant texts, his may be more reliable at the higher end than the naive estimate of p(B|A) that this macro produces.

The uses of association matrices are basically twofold at the exploratory level. They can be inspected, ideally together with the overall frequencies of the terms in the corpus, in table form. In this example, the p(B|A) are computed for the terms prominent in the LWI corpus of 76 documents. Excel's conditional formatting has been used to mark different ranges. Diagonal elements have been set to zero, for graphing purposes, an option that the macro offers.

Table 7: Association matrix for ten terms in the LWI corpus

	Target term										
Base term	*theolog*	*bless*	*critic*	*focus*	*importan*	*transit*	*secur*	*govern*	*academ*	*tragic*	**Occurrences in corpus**
theolog	0.00	0.14	0.21	0.46	0.43	0.07	0.07	0.46	0.07	0.07	100
bless	0.57	0.00	0.57	0.43	0.71	0.29	0.29	0.57	0.29	0.29	7
critic	0.55	0.36	0.00	0.64	0.64	0.18	0.36	0.82	0.27	0.18	16
focus	0.59	0.14	0.32	0.00	0.55	0.09	0.27	0.64	0.09	0.09	32
importan	0.67	0.28	0.39	0.67	0.00	0.11	0.22	0.56	0.11	0.17	29
transit	0.50	0.50	0.50	0.50	0.50	0.00	1.00	0.75	0.75	0.75	5
secur	0.15	0.15	0.31	0.46	0.31	0.31	0.00	0.69	0.31	0.23	30
govern	0.33	0.10	0.23	0.36	0.26	0.08	0.23	0.00	0.10	0.10	99
academ	0.50	0.50	0.75	0.50	0.50	0.75	1.00	1.00	0.00	0.75	5
tragic	0.40	0.40	0.40	0.40	0.60	0.60	0.60	0.80	0.60	0.00	5

Orange: >= 0.75
Yellow: 0.75 > x >= 0.5
Diagonal elements set to zero

Second, such a matrix can be turned into a network graph, which is a popular thing to do in this age of network sociology (e.g., Davies 2005: for the monitoring and evaluation of development projects), but harbors its own dangers. The network version of this table is given in Figure 8 on page 40, followed by some critical reservations.

[12] I did not implement his measure *ass*(A→B) because its interpretation is less than straightforward, with *ass*(A→A) < p(A|A) = 1. But it is from his work that I recognized the case for an asymmetrical measure. van Atteveldt's book is available form http://vanatteveldt.com/dissertation and makes for very useful reading for those venturing further up the learning curve into text analytics proper.

Extracting from internally stored text

Excel 2007: Up to 32,700 characters visible in one cell

The 1,024 displayed-character limitation on cells was released in Excel 2007; the new limit is over 32,000 characters or several pages' worth of a text document. Also, with more than 256 columns available in a sheet, large term co-occurrence matrices can be built, analyzed and visualized in network graphs. This opens new avenues for text analysis in Excel.

At the most basic level, we can now be sure that enough text can be stored and displayed in one cell in order to accommodate full sentences and (with unlikely exceptions) full paragraphs. These are natural units above the word and elementary semantic unit (such as the subject-verb-object triplet) level. Frequency and association patterns can now be analyzed at various levels of corpus decomposition - for entire documents, for paragraphs, for sentences or, if we develop suitable tools for text parsing, even for custom-tailored segments. Any text shorter than the limit can be copied and pasted entire into one cell. This allows us to access the texts of interest within Excel, without making calls to other applications.

The expression "paste text into spreadsheet cells" is somewhat disingenuous. An amount of text preparation may be needed in order to obtain the desired cell fillings. The segment to go into a particular cell needs to be ridden of paragraph marks and, ideally, also of other special characters such as tab marks or bullets. If the source document is in html or pdf format, it may be advisable to copy it in its entirety and paste it into a word processor, using the "Paste special (Unformatted text)" command.

Thus, with analysis inside Excel 2007, the challenge shifts. Instead of learning and tuning the word frequency macro discussed in the Excel 2003 context, the analyst now wants to import entire texts most efficiently, and in the types of segments that his objectives favor (sentences, paragraphs, entire texts).

Is there a universal algorithm for this purpose? Theoretically, yes, but the pecularities of text documents to be imported into spreadsheet cells vary. For repetitive tasks, one may want to record a sequence of appropriate commands (e.g., deletion of special characters, replacement of all or only of line-end para marks) in a Microsoft Word macro. A blank open Word documents may serve as a transit station between the source documents (in Word, Open Office, pdf, htm formats) and the recipient spreadsheet cells.

The following screenshot illustrates the results from a copy-paste process via text cleaning in Word. This is a segment from the 81-para (including title lines, excluding contact information) situation report (Sitrep #23) that the United Nations Office for the Coordination of Humanitarian Affairs released on February 22, 2010 regarding the status of the Haiti earthquake response. The original was downloaded in pdf format, then copied and pasted (unformatted text) into Word, where end-line para marks were replaced, and genuine para marks were retained. In Word, the entire cleaned text was selected. It was pasted into this spreadsheet, with cell R2C4 active. Excel then interpreted each paragraph

as wanting to go to a different cell. It flushed the text downward, into R2C4:R82C4. There was no need to count the paragraphs in Word, or to pre-select a range with exactly as many rows as there were paragraphs to accommodate.

Figure 8: Pasting text into Excel so that each paragraph is held in a separate cell

	1	2	3	4	5
1	**SitrepNo**	**Date**	**Para**	**Text**	**Length**
2	23	22-Feb-10	1	I. HIGHLIGHTS/KEY PRIORITIES	28
3	23	22-Feb-10	2	The Direction for Civil Protection (DCP) estimates that 222,517 people died following the 12 January earthquake, an increase of 5,000 people since the last estimate given a week ago.	182
4	23	22-Feb-10	3	The most urgent priorities for assistance continue to include shelter and sanitation.	85
5	23	22-Feb-10	4	There is also a critical need for rubble removal as well as for the identification of suitable land for the construction of transitional shelter. This is a major challenge for the decongestion of overcrowded sites.	214
6	23	22-Feb-10	5	The Ministry of Education has indicated that children in affected areas should resume school by early April.	108

Any variables subsequently calculated on paragraph text - as an example, the function LEN (length) was used to return the number of characters in the text cell - can be aggregated for the entire documents. A Pivot table using a unique document identifier - SitrepNo suggests itself - will do this. Evidently, cells sequentially filled with one paragraph each are a convenient way of accommodating texts longer than 32,700 characters.

For analyses of concept distributions, the approach already demonstrated for Excel 2003 (pages 2225 - 25) will work as well, using the same term flagging formula detailed in the appendix (page 48). This formula has been used for the table below. It returns, for each of 23 Haiti situation reports, the presence (1) or absence (0) of terms signifying particular response challenges. The examples chosen (out of about 250 initial terms) are of two first-phase needs (in red) and of two prominent slightly later. One notes, for example, that child protection had become a permanent concern of the responder community already before the attempted child abduction scandal broke on February 1, 2010 (Associated Press 2010).

Table 8: Examples of response challenges in 23 UNOCHA Haiti sitreps

SitrepNo	Date	Text	Length	search-and-rescue	water purification	child protection	cash-for-work
1	12-Jan-10	Haiti Earth	25,906	0	0	1	1
2	13-Jan-10	Haiti Earth	9,249	0	0	0	0
3	14-Jan-10	Haiti Earth	13,619	1	1	0	0
4	15-Jan-10	Haiti Earth	17,250	1	1	0	0
5	16-Jan-10	Haiti Earth	20,807	0	1	0	0
6	17-Jan-10	Haiti Earth	19,955	0	0	1	0
7	18-Jan-10	Haiti Earth	18,464	1	0	1	1
8	19-Jan-10	Haiti Earth	21,812	1	1	1	0
9	20-Jan-10	Haiti Earth	17,906	0	1	1	0
10	21-Jan-10	Haiti Earth	19,668	0	1	1	0
11	22-Jan-10	Haiti Earth	19,164	0	0	1	1
12	24-Jan-10	Haiti Earth	20,799	0	1	1	0
13	25-Jan-10	Haiti Earth	18,343	0	0	1	0
14	27-Jan-10	Haiti Earth	22,062	0	0	1	1
15	29-Jan-10	Haiti Earth	21,530	0	0	1	1
16	1-Feb-10	Haiti Earth	27,549	0	0	1	1
17	3-Feb-10	Haiti Earth	26,666	0	0	1	1
18	5-Feb-10	Haiti Earth	27,660	0	0	1	0
19	8-Feb-10	Haiti Earth	25,220	0	0	1	0
20	11-Feb-10	Haiti Earth	26,644	0	0	1	1
21	16-Feb-10	Haiti Earth	27,777	0	0	1	1
22	19-Feb-10	Haiti Earth	22,736	0	0	1	1
23	22-Feb-10	Haiti Earth	26,239	0	0	1	1

In this case, the entire text of each sitrep fills one cell in the third column ("Text"). The formulas used in column 5 and following are the term flagging formulas discussed earlier. As always, the choice of unit (entire text, paragraph, sentence, or other) and the use of appropriate formulas depend on the analytic objectives. The most useful formulas will likely be compounds of Excel text and lookup functions[13]. For example, term frequency within the text stored in a cell can be computed by exploiting the difference in text length after hypothetically deleting all instances. A combination of LEN and SUBSTITUTE, as in "=(LEN(*textRef*) -LEN(SUBSTITUTE(*textRef*,*termRef*,""))) / LEN(*termRef*)", might do the job. Such results then provide the raw material for verbal summary, further statistical analysis or graphic visualization, as needed.

As always, creativity and vigilance need to go hand in hand. A vigilant analyst will remember that, in the above sample formula, SUBSTITUTE is case-sensitive, demanding adjustments of the formula where terms occur in different cases. A creative member of an

[13] McRitchie (McRitchie 2008) is a good starting point for compound Excel function searches.

evaluation team, tasked to retrace the dynamics of a relief action, may want to throw numerous potentially relevant terms into the right-hand side of the spreadsheet holding the text corpus, copy the occurrence formula to the entire intersection of terms and records, then explore co-occurrences of interest by way of Pivot tables. A double-click on the Pivot table field intersection, e.g., "World Vision" and "cash-for-work", will produce a sub-table of all records in which both occur. Excel's Pivot table facilities thus produce concordances of the kind that TextSTAT delivers, enriched with all associated variables.

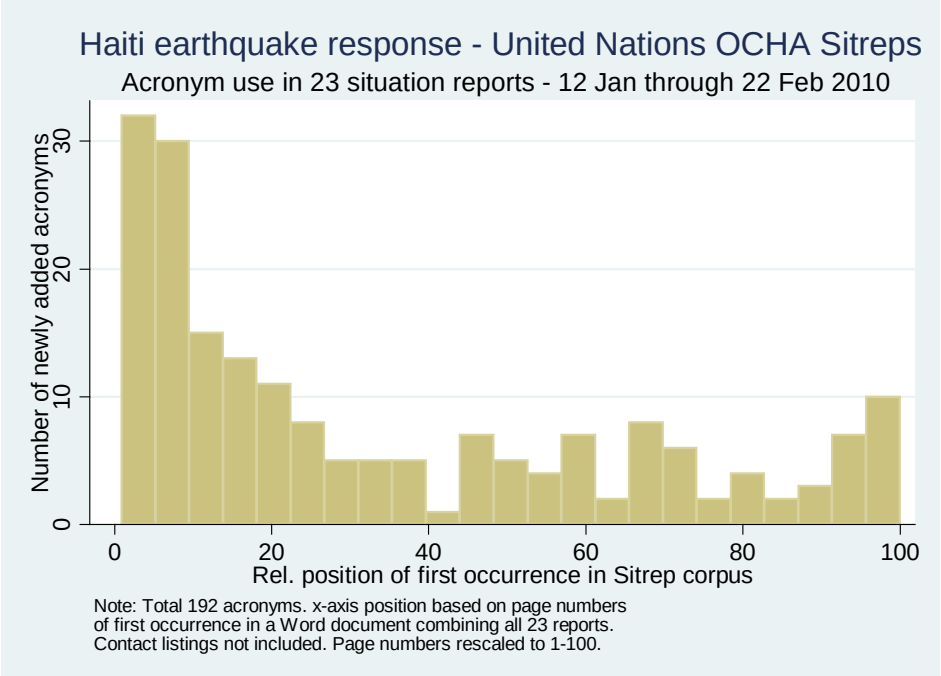
Analysis in Excel thus leads from text to numbers, and from numbers back to select text.

[Sidebar:] Network analysis within Excel

With Excel 2007, an open-source add-in for mid-level network analysis and visualization has come into the freeware marketplace. NodeXL (Smith and et.al. 2010) may not have all the analytic bells and whistles that mainstay network programs such as Pajek boast, but it is integrated with Excel, easy and intuitive to learn, and capable of turning matrices with large numbers of nodes and vertices into network diagrams. While it is a general-purpose network tool, it is suitable also to assist text analysis under time pressure. It grows increasingly productive, compared to manual diagramming with the Draw tools, as the number of elements escalates, or when the nodes of the graph have to be plotted efficiently, i.e. semi-automatically.

I illustrate its usefulness again with part of the Haiti sitrep material. One of the challenges in program assessment is to keep on top of an ever growing list of actors, their acronyms as well as the activity lines and issues with which they are connected. In Haiti, within six weeks, the OCHA sitreps made use of nearly 200 distinct acronyms. Most of them were introduced in the first two weeks, as the first-occurrence histogram below demonstrates, with a fluctuating trickle of new arrivals later on. This pattern may be typical of an organizational network with high levels of pre-established readiness and mutual acquaintance.

Figure 9: Acronym introductions in the UNOCHA Haiti sitreps

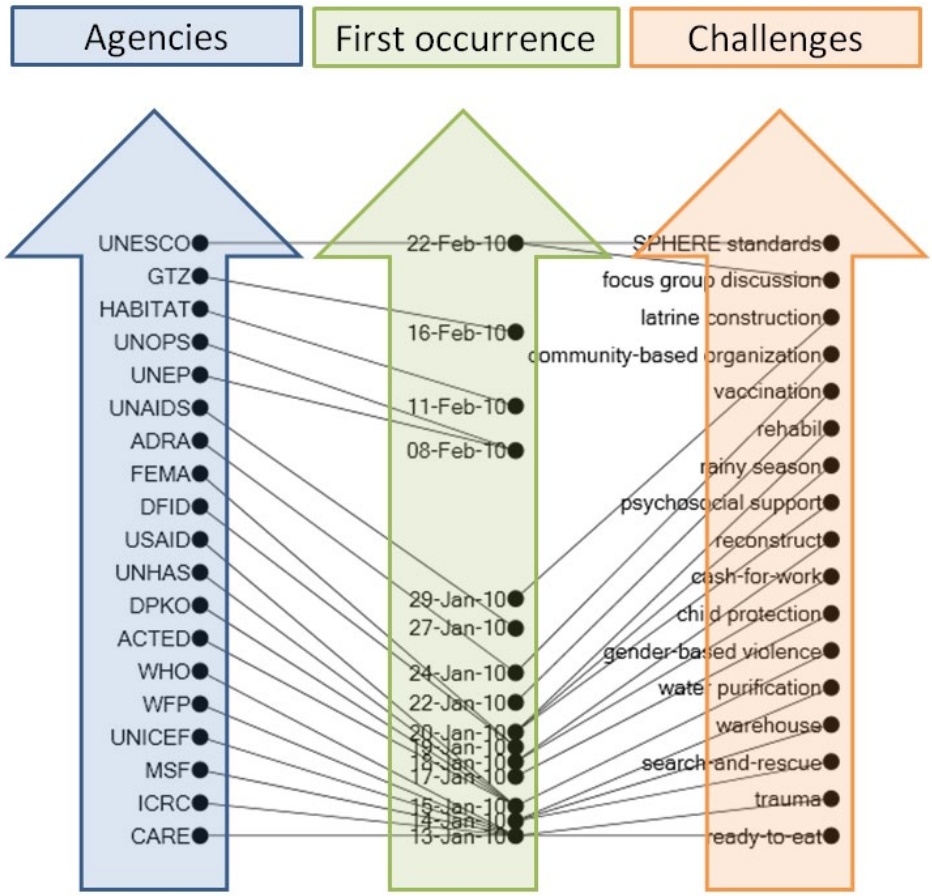


Confronted with a plethora of actors and, from within the sitrep texts, programmatic terminology, the analyst may easily be overwhelmed by the challenge to reconstruct clear, graphic timelines. Text functions and NodeXL may come to the rescue:

- Once the texts have been suitably lodged in the spreadsheet, term occurrences can be calculated for each record. These records have approximate date or time stamps, such as through the release dates of the reports in which they appeared. The date when a term of interest (agency name, acronym, issue or programmatic term, etc.) first occurred can be extracted using a combination of Excel functions (such as MATCH and INDEX).

- Pairs of dates and entities of interest (agencies, issues) then form the raw material for a NodeXL graph. In the sample graph below, selections of relief agencies and challenges are ordered sequentially by first appearance in sitreps, and are tied by lines to those dates.

Figure 10: NodeXL graph of Haiti sitrep terms linked to dates of first occurrence



Note: The first OCHA sitrep, of 12 January 2010, was not included in the calculation of first occurrences because its character was more of an alert than of agency activity.

This picture is crude and selective. It reduces over 100 actors (most of the acronyms stand for agencies) and over 200 salient multi-word terms (abstracted from the sitrep corpus) to a small number fit for visualization.

Nevertheless, it does illustrate some of the response dynamic. NGOs with a long-standing presence in Haiti (e.g. CARE) were among the earliest responders. Others, with a rehabilitation and development mandate (Habitat, GTZ), would come to sitrep attention much later. "Ready to eat" meals were important in the first response; reliance on community-based organizations became a topic after two weeks; focus group discussions must have been considered a luxury item in the relief tool box until after several more weeks.

The point here is the tool: With NodeXL, texts describing complex actions can rapidly be converted via intermediary numeric data into network graphs, all of this within Excel.

Tool #3: Rapid extraction with STATA's Wordscores

We move one level up on the depth-of-analysis ladder by placing text analysis within full-fledged statistical applications. Such programs allow the analyst to unearth hidden relationship in the text structure in greater variety, speed, and analytic validity than what Excel offers. The trade-offs, however, are considerable; they include cost – STATA, for example, is a commercial program -, loss of participation by team members and local counterparts not familiar with the application, and suspicion or even outright rejection by stake-holders of the use, or results, of models that appear exotic and uncontrollable. When taking text analysis to such levels, one has to make a circumspect decision considering both likely benefits and costs.

Nevertheless, statistical analysis is widely used also in policy studies, evaluation and monitoring. STATA (StataCorp 2007) frequently is the instrument of choice in survey analyses. Familiarity with it has been noted as a requirement in some evaluation ToR. If I present a text analysis tool written for STATA, this is justifiable also on the grounds that freeware statistical packages with text analysis modules exist as well[14].

Wordscores is a small STATA package originally written for the analysis of political party programs (Lowe 2008; Benoit, Laver et al. Undated). It is ideal for producing large word frequency tables using all the one-string terms it finds in the corpus. Its other attractions are its rapid speed and its ability to stem word forms. "Stemming" is a text preprocessing operation that removes common morphological and inflectional endings from words, reducing them to the radicals that bear the essential meaning. Stemming thus merges different forms of the same word (or of closely related words) that should be treated as one in the analysis. Thus, plural and singular would be reduced to one form: "concern" and "concerns", but also "concerning" and "concerned".

The package contains several procedures, among which two are of interest here:

[14] The most widely used statistical freeware, R, is probably also the most diversified in the choice of specialized modules ("libraries"). The text analysis features that STATA's Wordscores offers, and many more, are available in its *tm* library (Feinerer, Hornik et al. 2008; Feinerer and Wild 2009). This is part of a much more ambitious research program than the level with which we are dealing here. But: All is free.

wordfreqj reads in a set of text files and

> *"produces a set of frequencies of all words that occur in at least one of the input texts. The resulting STATA dataset consists of a text variable word containing a list of the words themselves, followed by a set of frequency variables, one for each text, with the names tfilename1, tfilename2, tfilename3, etc. Each frequency variable will range from 0 to a maximum of the total words associated with its text file"* (Benoit et al., op.cit.).

Stemming is optional.

The total occurrences in the corpus have to be computed by the user; I inserted this statistic ("OccurCorpus"), plus the proportionate relative frequencies ("PerMillionWords") in this partial output. The segment shown here displays the frequencies for the 11 most often used words in the first three articles of the LWI corpus (see page 26).

Table 9: A segment of a frequency table extracted with STATA's Wordscores

Word	OccurCorpus	PerMillionWords	tlwi2008001	tlwi2008002	tlwi2008003
the	3423	79472	20	37	36
of	1576	36590	14	14	38
and	1509	35034	8	27	22
in	1388	32225	5	21	19
to	1151	26723	9	14	20
church	758	17598	1	18	7
a	683	15857	4	5	5
lwf	584	13559	5	12	7
for	563	13071	3	6	7
on	438	10169	0	4	7
lutheran	410	9519	1	9	3

The second procedure of interest is called ***describetext***. It lists, for each text in the analyzed corpus, the total word count, the number of unique words, and the mean and median frequency of the word use. By first calculating the word frequencies in the corpus ("OccurCorpus" in the table[15]), one can ask *describetext* to return also the total number of unique words in the corpus. This, by design, is equal to the number of records in the frequency table. For our LWF corpus, *describetext* reports 3,726 unique words in 43,072 words total. This is more parsimonious than the 6,232 word forms / types which TextSTAT reports for 44,577 words / tokens. The economy is the result of stemming and excluding numbers. A segment of the resulting table looks like this:

[15] Using *egen(OccurCorpus) = rsum(tlwi2008*).*

Table 10: Summary statistics for each document in Wordscores

Text	Total Words	Unique Words	Mean Freq.	Median Freq.
tlwi2008001	297	154	1.93	1
tlwi2008002	534	233	2.29	1
tlwi2008003	582	215	2.71	1
tlwi2008004	97	57	1.7	1
tlwi2008005	1,250	505	2.48	1
tlwi2008006	1,021	366	2.79	1
tlwi2008007	1,415	252	5.62	1
[etc.]				
freqTotal	43,072	3,726	11.56	2

The article LWI2008-007 (which Wordscores returns as "tlwi2008007") stands out for its unusually parsimonious vocabulary or, which is the same, high mean word frequency (5.62). As the reader may recall, this piece reported on church growth in the Lutheran community, using the words "Lutheran" and "church" with rare high frequency. *describetext* can thus be used to flag members of the corpus with likely atypical topics and vocabulary.

What does this all yield? For an example of a small analysis piece using Wordscores, this sidebar looks at the coherence of the 76 LWI news articles that I already introduced as an experimental corpus.

[Sidebar:] The LWI corpus – Dimensionality and semantic network

Which among the particularly prominent words in this corpus tend to appear together in the articles? What kind of preoccupations or foci do these clusters, if there are any, suggest?

Obviously, that depends on the selection of the terms on which we compare the 76 articles. As in many other English language texts, grammatical particles like "the", "of", and "and" are the most frequent words. Except for questions of style, they do not make suitable comparators. Similarly, the self-referent "Lutheran" and acronyms like "LWF" are too common to provide meaningful discriminant markers for most conceivable analyses.

Unless meaningful comparators are already known from research perspectives or from pre-existing acquaintance, they have to be identified. Terms may be pre-selected if they occur in the corpus significantly more frequently than in some other corpus that the researcher considers an interesting comparison base. For example, Landmann and Zuell (2008), in attempting to identify contemporary public events in their corpus of interest, created a large collection of texts from the British newspaper "The Guardian". This was supposed to represent general language usage.

I adapted their four-step analysis procedure (op.cit.: 488) to flag prominent terms in the LWF corpus:

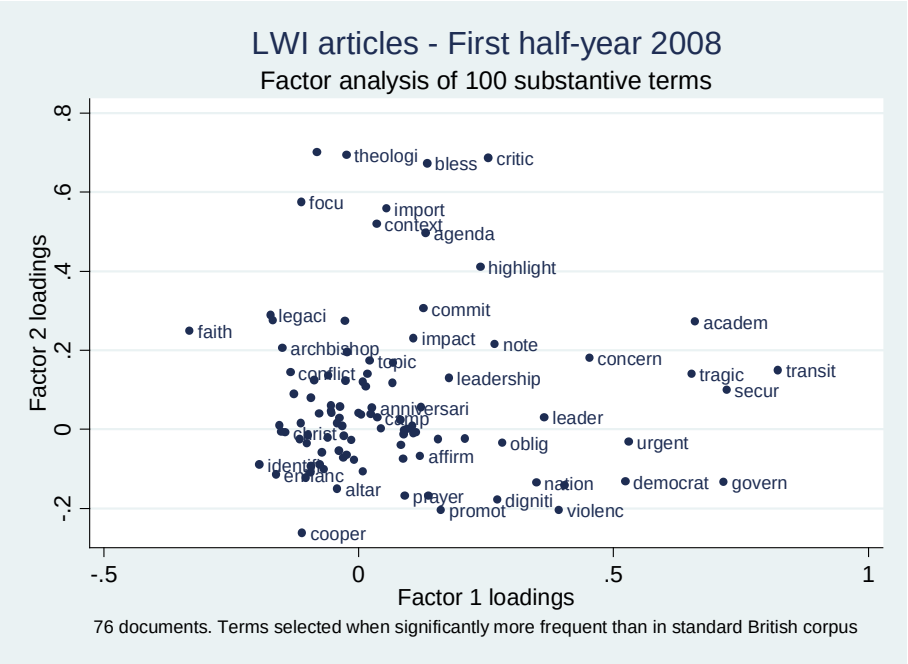
- Step 1: Acquisition of the corpus of interest: Download and save all LWI articles for the period of time concerned
- Step 2: Acquisition of the reference-text corpus that represents general language usage
- Step 3: Word selection:

- a.: Calculation of word frequencies and of relative frequencies for all words in both corpora
- b.: Calculation of differences between the relative word frequencies of specific words in the reference-text corpus vs. event-text
- c.: Selection of the *n* words with the largest differences between general language usage and the event texts

- Step 4: Exploratory factor analysis based on the selected words.

For reference, I used the companion frequency list to *"Word frequencies in written and spoken English: based on the British National Corpus"* (Leech, Rayson et al. 2001b; Leech, Rayson et al. 2001a). This had the advantage that the frequencies for the British corpus were already calculated. After stemming, its terms were matched to the LWF word list[16]. From a list of 3,174 matches, words were pre-selected if they occurred in both corpora more than 5 times per 1 million words and if the relative frequency in the LWF corpus was more than 5 times greater than in the British corpus. This yielded a subset of 282 terms. For flavor, a few examples may suffice: Lutheran – 4,758 times more frequent in the LWF corpus; children – 672 times; Rwanda – 556 times; divers[-e, -ity] – 254 times. I whittled the 282 down to 100 manually, eliminating geographical and most other named entities as well as ambiguous terms.

Figure 11: Factor analysis of 100 terms in the LWI corpus



For this sidebar, I abbreviated the process, presenting a partial result only. By transposing the term-document matrix, the 100 key terms became the variables which I factor-analyzed in the final step of the Landmann-Zuell procedure. The graph on the previous page displays the "clustering" of the terms along two independence axes.

Two observations impose themselves:

- Compared to the *outside* (everyday language) corpus, there is a remarkable prominence of emphatic and appellative terms such as "note", "commit", "affirm", "focus", "import[ant]",

[16] This process was more convoluted than this sentence admits and defies brief description.

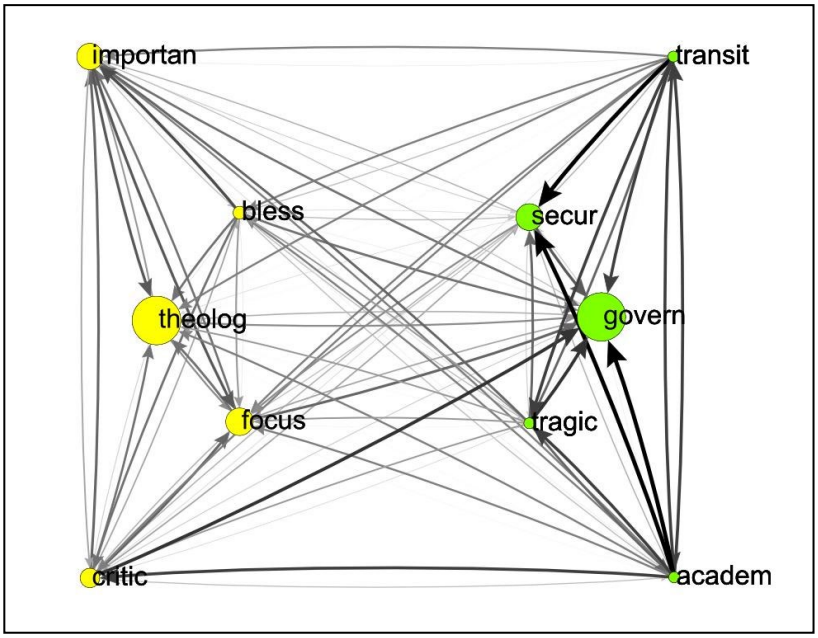
"critic[al]". This may be congenial to the organizational form of the Lutheran World Federation, a federation in which member consensus is crucial, and which engages in advocacy work on many fronts. It may also result from a particular editorial style in which the content side of the information is carefully managed together with selectivity of the communication, in order to ensure attention and acceptance (Luhmann 1987: 196).

- What distinguishes the articles *within* the corpus? The first factor exhibits a cluster (on the lower right-hand side) that is more action and governance oriented (e.g., govern*, secure*, democrat*, violenc*). It is distinct from the obverse cluster (left side) that appears to express identity concerns (e.g., faith, identific*, Christ [incl. Christian], legaci*, anniversari*). The second factor – see the cluster in the top area – may highlight the explicit theological work that the Federation facilitates (e.g., theologi*, focu*, import*, context, agenda, bless*). The interpretation of the obverse cluster (see bottom points) is not straightforward – one would not want to assume that people gather around the altar for prayer only when theologians are not around!

In fact, these differences are mild when we look at the total factor structure. The leading factors explain a small portion of the occurrence variance only. The first factor accounts for 7.5 percent, and it takes 11 factors to absorb half of the variance between the 76 articles. This indicates that the particular institutional vocabulary is used in a fairly integrated way across topics and authors. There are no major semantic cleavages *in this particular set* of public articles.

Factor loadings are a familiar concept for a minority of professionals who analyze data in the humanitarian and development fields and may remain ever foreign to their consumers. It may therefore be helpful to present some of the factor analysis insights in visual aids that are more commonly intuitive. What maps are to statistics, networks may be to statistical charts – a partial representation of, or a different approach to, an overly complex or opaque set of relationships. For example, for the five concepts with the highest loadings on either of the first two factors, I constructed this semantic network representation based on an association matrix. The matrix was calculated with the help of an Excel macro described in the main body and fully given in the appendix.

Figure 12: Semantic network representation of 10 terms in the LWI corpus



The more theological concepts (yellow) are arranged on the left side, and the more governance-oriented (green) on the right, with vertex areas expressing frequency in the LWI corpus. An association pattern appears distinctly. Associations within the governance cluster seem, by and large, to be stronger than those among the five terms of the theological cluster. For example, a majority of LWI news pieces that carried the concept *transit* (transit, transition, etc.)* would make use also of the concept *secure* (secure, security, etc.)*. Between clusters, strong associations run from *academ** to *critic**, and hence back to *govern**. Associations are not symmetrical; only a very thin arc leads back from *govern** to *critic**.

However didactically appealing the network representation may be, there are two downsides. As one can readily see in this graph, most of the strong and very strong associations originate from less frequently used concepts (*"transition"* occurs 5 times in the corpus, *govern** 99 times). This calls into question the sampling stability of these claimed strong semantic relationships. Second, didactic consideration again oblige us to limit the network graph to a small number of nodes (= concepts). By doing so, however, the analysis makes use of a small part of the information, compared to the wealth that goes into statistical procedures.

There is thus a trade-off between aesthetic appeal and intuition on one hand, and validity and reliability on the other. Ultimately, this can give rise to ethical dilemmas. A complacent analyst may feel that his client will not take ownership of statistically supported findings, but will be pleased with maps and network graphs that purportedly demonstrate some positive point – or at least insinuate a profound grasp of the project's reality! A more sincere analyst would qualify his findings with the necessary provisos, at the risk of losing the initial enthusiasm of principals who want straightforward messages.

Statistical text analysis thus leads fairly rapidly to interesting second-order questions, some of which can perhaps be elucidated by further, keener statistical intrusion. Most of the answers, however, should be expected of more classic forms of text analysis (which begin by actually reading the texts!) and from going to its authors or to the people that they introduced. Whether factor analysis or rather some other particular statistical procedure is appropriate for the material at hand, is not of concern here – the point is that Wordscores results can be quickly and powerfully analyzed. This output then can be used to stimulate discussion in an evaluation, research or other open-minded setting.

Climbing higher on the learning curve?

TextSTAT, text functions and a word-frequency extracting macro in Excel and STATA's Wordscores are the three steps at which I can assist the reader with some detailed explanations. Naturally, many will want to explore higher ranges of the learning curve, in hopes of acquiring ever stronger text analysis tools. This is where I have to part with the reader, out of sheer inexperience or in doubts about further productivity gains in the environment in which my target audience works. But I would like to share a couple impressions that I formed while peeking into some of these higher spheres, and also offer a sidebar on a set of operations that are routinely performed as entry toll to them. These concern data preparation.

Text analytics

With the expansion of the Internet, text analysis has been expanded – some might say: hyped up – to the new field of text analytics. This field, which Wikipedia defines as a

> *"set of linguistic, lexical, pattern recognition, extraction, tagging/structuring, visualization, and predictive techniques. The term also describes processes that apply these techniques, whether independently or in conjunction with query and analysis of fielded, numerical data, to solve business problems. These techniques and processes discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing"*[17] ,

is vigorously growing in alliances among academic, corporate, government and military sectors. From these, the NGO part of the aid industry, the social movements that they support as well as many academic researchers have habitually kept their distance. Text analytics helps organizations manage large databases with textual elements, in call centers, e-mail intercepting intelligence, bioinformatics, and technical libraries – tasks that the "assistance sector" of society often does not need, or not want, to undertake, or has allied organizations (e.g., social welfare bureaucracies and insurance companies) taking care of. Large international humanitarian and development NGOs may have begun to exploit related applications, such as in knowledge management, but I admittedly know little of this.

There are a number of open-source text analytic applications. The Natural Language Processing Group at Sheffield University, United Kingdom, has developed "GATE – General Architecture for Text Engineering"[18], which claims a large user community but comes with a modular structure that may confuse newcomers. RapidMiner Enterprise[19] offers a free community edition of its data mining software, but I have not been able to download the associated text analysis plug-in. AutoMap, a software developed by the Carnegie Mellon University, Center for Computational Analysis of Social and Organizational Systems (CASOS)[20] (Carley 2009), has good text preparation tools (see Sidebar below). Tutorials and sponsorship indicate that lately it has been enhanced chiefly for terrorist network detection.

This is a fast moving field with which the documentation, public upgrades and reasonable effort to make informed selections can barely keep pace. Without an education in linguistics, even the statistically minded outsider can participate only up to a point. For the hurried consultant (and for the harried monitor), these applications seem too heavy and too solitary in their community. An exception could be made for situations in which network structures – physical, social and semantic – have to be investigated deeply. An example that comes to mind is Moore et al.'s (2003) study of the Mozambique 2000 flood response, assuming that the authors collected a considerable number of documents from the 65 NGOs in the network.

[17] http://en.wikipedia.org/wiki/Text_analytics.
[18] http://gate.ac.uk/index.html
[19] http://rapid-i.com/component/option,com_frontpage/Itemid,1/lang,en/
[20] http://www.casos.cs.cmu.edu/projects/automap/ . Updated in June 2009, after my initial trial.

[Sidebar:] Text preprocessing for effective analysis

TextSTAT takes word forms as they come. The search string “empower”, for example, returns “empower”, “empowering”, and “empowerment”. Their instances must be inspected in a separate concordance for each form. This is not optimal in the search for concept prominence and for underlying meaning structures. As we have already seen in the Wordscores section, reducing words to their radicals, through an operation called stemming, may make for better and easier analysis. Stemming is one of several text preprocessing steps that natural language processing software such as AutoMap provides beyond the elementary functionalities of TextSTAT and Wordscores. A brief enumeration of these operations may give a first idea of the processes involved. I largely follow Leser (2008), with some additions from Carley (2009):

- **Format conversion**: The software may require conversion of all corpus documents into one particular format that it can read, such as .txt.
- **Removal of special characters and/or numbers**: This facilitates indexing and searching.
- **Conversion to lower case**: Combines words that happen to be lower or title case by accident of sentence position, but loses abbreviations and makes named entity recognition more difficult.
- **Stop word removal**: Frequent words whose removal does not normally change document meaning in text analytics (it does in everyday language, including our normal reading!). The ten most frequent stop words in English are: the, of, and, to, a, in, that, is, was, it. Removing the top six (the, of, and, to, a, in) typically eliminates a fifth of the tokens.
- **Named entity recognition**: Proper names are very important for the naive understanding, but also for the latent meaning search in texts. Many consist of more than one token. The Lutheran World Federation is not some federation of all Lutheran worlds, whatever this could mean, but the worldwide federation of Lutheran churches.
- **Speech tagging**: Attaching to each word a tag of its supposed position / function within its sentence later helps with processing the text in “actor – organization – activities” and similar schemes.
- **Anaphora resolution**: In natural language, the meaning of most pronouns is made clear by grammar and context. “The LWF delegates passed two resolutions. They discussed them again the following morning”, meaning “The delegates discussed the resolutions again..”. In text analytics, this may need to be made explicit.
- **Stemming**: Reduces words to their base forms so that different word forms with the same meaning are collapsed. Often these are neither a standard word in the language (e.g., “theolog”), nor the exact linguistic root.
- **Thesaurus creation**: A set of fixed terms and relationships between them allows texts to be organized in hierarchical manner. Thus “group liability” may be part of “loan repayment”, but not of “technical support”. Both may be part of “microcredit”.

To repeat, the intent is not to equip oneself for all these operations, but to acquire a sensibility for some of the ways in which modern text analytics deals with linguistic complexity.

Qualitative research software

Much of this paper has so far dealt with word lists or term lists. Common sense and linguistics, however, tell us that meaning resides in sentences rather than in words. In fact, it resides in word, sentence and wider context – in what preceded, and (if already known or anticipated) in what follows.

The necessity to pay close attention to meaning structures is one of numerous reasons that have spawned an explosion of qualitative research, and more recently also of "mixed methods" approaches (the combined use of qualitative and quantitative methods). The methodological field is vast and growing (Denzin and Lincoln (2005) is one among many large handbook-type works) and does not concern us here except to point to the existence of text analysis applications specifically couched in qualitative research traditions.

Apart from commercial packages, some of which have attracted a community of users of consequential size and support – leaders include "Atlas.ti" (Hwang 2008) and "QDA Miner" (Lewis and Maas 2007) -, a few open-source applications are available (for links to some, see again Altman, op.cit.). Notable for the institutional prominence of their sponsor, EZ-Text[21] and AnSWR[22] are two applications created within the US Center for Disease Control (CDC), primarily to support qualitative research with patients.

Given scant experience with such applications, I limit my observations to two, regarding both text analytics and qualitative research:

First, the learning curve is clearly much higher. Some of the research institutes or sellers behind such software organize training courses; typically these last a full week. As to freeware, more than once I found that the documentation was outdated (it taught an earlier version) or too abridged to guide self-learners through the initial hurdles.

Second, it is true that humanitarian and development evaluation ToR occasionally demand qualitative approaches. The flash word for these kinds of expectations is "triangulation". But one may wonder whether the desk officers drafting the ToR are conscious of the challenges that serious triangulation places on an evaluation team and its host organization. As far as the computer applications are concerned, some other factors conspire against their use in evaluations and similar assignments. Apart from rare and lucky partnerships with local academics already familiar with the particular program that the expatriate team member brings to the task, reliance on advanced software during team work may turn the user into a social and cognitive isolate.

Discussing barriers to successful mixed-method approaches, Bryman (2007) explicitly mentions synchronization issues: *"The timelines of the quantitative and qualitative components may be out of kilter so that one is completed sooner than the other"* (ibd.: 14). Which side advances faster depends also on institutional barriers to acquiring

[21] http://www.cdc.gov/hiv/topics/surveillance/resources/software/ez-text/index.htm

[22] http://www.cdc.gov/hiv/topics/surveillance/resources/software/answr/index.htm . I installed this software on a computer some years back, but found the documentation insufficient. The AnSWR Web page has not been modified after May 2007.

documents speedily, say, for example, policy documents from capital city headquarters vs. spreadsheets from decentralized field monitoring units[23]. Bamberger et al. (op.cit.: 84) are generally pessimistic about the use of qualitative data analysis packages under time pressure: *"they take a long time to set up and the purpose is usually to provide more comprehensive analysis rather than to save time."*

This does not preclude that situations exist in evaluations and field research in which advanced text analytical and qualitative research software significantly enhances productivity. Davis, in a workshop report on panel surveys and life history methods (Baulch and Scott 2006), relates the use of such a program for the subsequent categorization of life histories that he collected among the poor of Bangladesh (ibd.: 8). Yet, by and large, the decision to invest the time (and, for commercial products, money) in learning and working with such applications must be weighted by the individual researcher considering her personal situation.

[Sidebar:] Food vendors and meaning structures

[23] I have been embroiled in similar dilemmas myself. At one time, I was hired as the number cruncher in a politically sensitive review of a large UN humanitarian program (and creatively designated as "relief economist"). All the other researchers in the team were qualitative-leaning. Due to the accident of data acquisition, I was the only one with "showable output" by the time the team presented at a conference attended by openly hostile government bureaucrats. Predictably, the presentation of relief goods transportation scenarios was singled out for contextual gaps. These were caused by the delay in working up historical and institutional aspects. Research software played a minor role in this to the extent that the political sensitivity obliged my fellow team members to reference hosts of slowly arriving documents in time-consuming watertight bibliographical annotations.

Another young man in Dili, East Timor, let me take this picture of his pretty arrangement of clementines, exuding a tranquility free from all time pressures. Different from his age mates on the title page, he sells an article unqualified by any texts, carefully managed with a local technology.

Yet, tranquility is the exception, not the rule. Itinerant food vendors move almost continuously, rapidly commuting between places and hours that incline their customers to buy. The work is hard, competition stiff. The man bore two such clusters, dangling from a shoulder yoke.

The picture does drive home a point in *text analysis*. Shaped by the physical properties of fruit and string, as well as by the man's stamina, marketing savvy and personal preferences, the cluster behaves as an analog computer. We notice the hexagonal compaction; the position of every fruit can be described with just a few parameters of an almost perfect lattice.

At the same time, this high degree of order conveys no knowledge whatsoever of the properties of other emergent levels. Seeing the cluster tells us nothing about whether these fruit have seeds, or how much money the vendor makes when he sells them. Similarly, the statistical structures that text analysis may detect say nothing about the ultimate meaning of a text as a whole, let alone of its pragmatic consequences. They do give us internal landmarks that facilitate the holistic quest.

Outlook: The dictatorship of time and the community of learners

This paper made four basic assumptions:

1. Humanitarian and development workers at times work with voluminous, complex or otherwise difficult text documents.
2. Such situations may necessitate more than revision, ultimately prompting a new text that interprets those at hand.
3. Often this type of work needs to be done within tough time constraints.
4. Computer-assisted text analysis can make it more efficient.

The temporal dimension is thus the leading one in this rationale. This can be questioned. The social and substantive dimensions of working with the text documents of relief agencies, social movement NGOs, the Red Cross, etc. may seem, in the minds of some, to hold more important directives. After all, what follows from the fact of life that time is always short?

The social dimension covers the reliability of text analysis – would another consultant interpret the same texts differently? – as well as such other aspects as the impact of the digital divide on collaborative arrangements. In the substantive dimension, there are validity challenges. It is not unknown to find reports, some with far-reaching claims, in which "text analysis" is hardly more than a codeword for insufficient field exposure. And, do the constructs and metrics of text analysis actually prove anything beyond, or distinct from, what the original texts purport to convey?

These are important questions, but I defend the "dictatorship of time" on two grounds:

- Any intelligent reading of texts is time-consuming. The discussion and synthesis, in working teams and then with principals, of the findings may take even more time. Devices that accelerate the initial processing of texts liberate time for later synthesis, debate and other important activities such as field visits. They help to redistribute the elements of learning processes while at the same time giving us a firmer handle on those texts of which we must take note.

- Second, besides the chronological and social time of the group that works with a shared set of texts, every participant lives his and her own biographic time. This includes the rhythms at which we replenish our professional and technical skills. You and I lose some skills inadvertently, shed obsolete ones deliberately, strive for some beneficial new ones, and remain ignorant of many others that would pay even greater dividends. We don't do it alone. Yet, the windows for learning together, across social boundaries and divergent agendas, remain open for brief moments only. Alone I learn for years; this particular group together - maybe for one hour. If others are to use my tools, I need to arrange a rapid transfer.

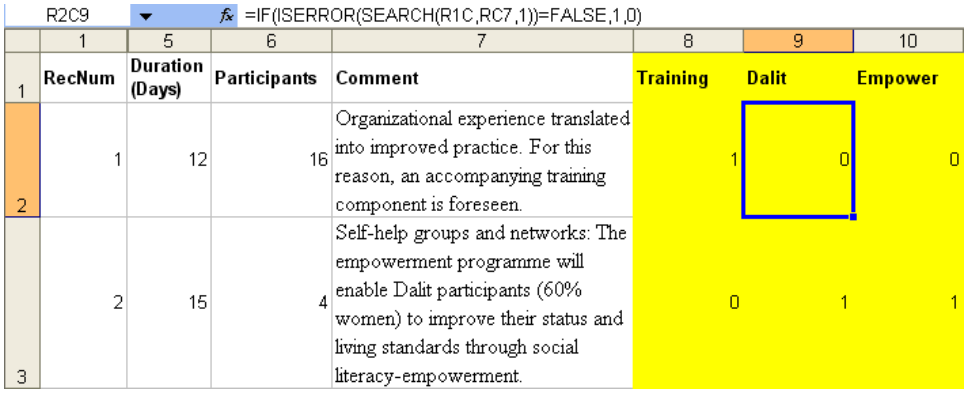
We must not wax philosophical. Much of contemporary learning is prompted and structured by developments of computing, down to the advent of a new piece of software, or the different use of an existing one. This is the situation that frames text analysis also in the humanitarian and development fields. The tools offered here, if they are any good, will speed up our understanding of texts, yet not leave it less profound.

Appendices

Excel search term flagging formulas

Step 1 of the three-step procedure outlined on pages 23 sqq. requires the creation of binary variables to indicate ("flag") the presence or not, in the concerned text field ("Comment"), of selected search terms of interest. For each term, a separate binary is created; the term is the field name and thus bolded in the top row. This screenshot image holds the same information of interest as the table in the main body, plus (yellow) the area of indicator variables as well as the formula bar, with the formula used in the active cell R2C9 (cell I2 in "A1"-notation).

Figure 13: Screenshot of a log book-like table, with search term indicator variables



R2C9 =IF(ISERROR(SEARCH(R1C,RC7,1))=FALSE,1,0)

	1	5	6	7	8	9	10
1	**RecNum**	**Duration (Days)**	**Participants**	**Comment**	**Training**	**Dalit**	**Empower**
2	1	12	16	Organizational experience translated into improved practice. For this reason, an accompanying training component is foreseen.	1	0	0
3	2	15	4	Self-help groups and networks: The empowerment programme will enable Dalit participants (60% women) to improve their status and living standards through social literacy-empowerment.	0	1	1

In this example, the three questions of interest are – whether the event was a training event, whether the Dalit community was mentioned (as participants, or in other situations), and whether a reference to the empowerment concept was made. The corresponding indicator variables take the value "1" if the term occurred in the comment, and "0" if it was absent. The results are shown in the yellow area.

The composite formula

=IF(ISERROR(SEARCH(R1C,RC7,1))=FALSE,1,0)

is used to calculate the indicator values. At first sight, it looks daunting. Yet it simply exploits a combination of text search, error reporting and IF functions. As an added bonus, mixed cell references are used to make the formula identically usable for alls search term fields (yellow columns) and event records (rows). The clarity of mixed references in R1C1-style is one of the reasons why this notation is preferable.

We break the formula open starting from the inside. In this example, we use a comment for which the result is 0, i.e. the term is not present. The logic is the same for terms present, as the dissection of the formula below will make clear. Novice readers may want to first study the explanations that Excel's Help function offers for the functions involved.

R2C9 is the active cell (blue border). It returns "0" because the search term "Dalit" does not occur in the comment field of record no. 1. How did we get this result?

Table 11: Deconstructing the term flagging formula

Functions used in formula:	How it works:	Formula result:
=SEARCH(R1C,RC7,1)	The reference R1C in "=SEARCH" directs Excel to copy the search term in row 1 and in the same column as the active cell. Here, this is "Dalit". The second argument, RC7, fetches the text to be searched, from the same row as the active cell, and in column 7. The third argument ("1") instructs to begin the search with the first character of the text to be searched. However, in the targeted cell R2C7 "Dalit" does not occur. "SEARCH" cannot find the desired term and therefore sends the error message "#VALUE!" [If it found it, it would send a number (the position of the search term within the text), not an error] We embrace "SEARCH" in the function "=ISERROR", to tell us if this indeed is an error:	#VALUE!
=ISERROR(SEARCH(R1C,RC7,1))	It returns "TRUE", "Yes, this is an error" Now that we know whether the search term occurs in the text or not, we add another function, "=IF". This formula tells Excel to return "1" if there is no error, i.e. the term occurs somewhere in the text. It returns "0" if there is an error, i.e., the text does not contain any of the strings "Dalit" or "Dalits" or "dalit".	TRUE
=IF(ISERROR(SEARCH(R1C,RC7,1))=FALSE,1,0)	"=IF(ISERROR(SEARCH(R1C,RC7,1))=TRUE,0,1)" does the same and may be more cogent in the minds of some.	0

The formula is convenient for two reasons. First, no matter how many search terms are of interest, and how many event records the table holds, the formula can be copied and pasted identical to the entire target range. Second, if a search term is changed (say, instead of searching for "Dalit" you wish to find occurrences of "women"), the results will automatically be updated. There is no need to change any formulas; just replace the search term in the cell of the top row[24].

Excel macros

Term frequencies in external documents

This macro (for MS Word documents) and the following (for .txt documents) require MS Word to be installed. They can be placed – by copying the entire text here in Courier New font – in any one of three different locations:

[24] Haynes (2009), using a similar formula, reports an application from an accounting department that used a text field to record reasons for delayed payments. Three columns were used to flag the most common reasons, a forth for prompt payments. These flags helped to reduce the initial set of 300,000 payment records to the roughly 100,000 that were of special interest to a research project.

- By right-clicking the sheet name tab of the worksheet that is to hold the frequencies, then clicking “View code”, the Excel VBA editor opens a space for macros and functions reserved to affect this particular worksheet. If stored here, the macros can be used only in this sheet.
- By inserting a module, in the VBA editor, for this workbook, or adding the code to an existing module, the macros will be available to work in any worksheets of any open workbook as long as the macro-holding workbook stays open.
- Some users may have an invisible template workbook that loads at Excel startup, most often named “Personal.xls”. When macros are placed in a module here, they will be available to all open workbooks any time and in all worksheets.

Here starts the macro code:

```
Sub FreqGivenTermsMSWorddocs()

'Excel macro to calculate frequency of search string in several Word documents
'Search strings to use must be listed in the leftmost column of the active sheet,
'starting in row 2. (Cell R1C1 or A1 can be used as a title "SearchTerms")
'The macro writes the document names in the top row, starting in R1C2 (or "B2").
'In the row below the word frequency matrix, it writes the total word count
'for each document.

'Example of a segment of the resulting table:

'searchstrings   test1.doc   test2.doc
'xxx 2   2
'yy  1   7
'aa  0   1
'xxxx    1   1
'TotalWordsInDoc 4   6

' where 7 means that the string "yy" occurs 7 times in document test2.doc.


'Written by Aldo Benini, 5 June 2009
'In part adapted from a macro written by Ray Tweedale,
'"Import WILDCARD search to EXCEL from WORD", 2008

Dim sPth As String 'Path where Word docs are stored
Dim sNam As String 'Name of Word doc file
Dim oWrd As Object 'MS Word as an application object within Excel

'-----------------------

'PARAMETERS THAT YOU MUST SET:

'1. Subdirectory where Word documents to investigate are stored:
sPth = "C:\Aldo\A_Automap\counttest\"
'Change as appropriate. Ensure backslash at the end.

'2. Column 1 (or "A" depending on cell reference type) holds the search terms
' from row 2 to the last row, without gap. Set the row number of the bottom cell
' (the one holding the last term in your list):
RowNoBottomList = 5 'Change as appropriate

'-----------------------
'Count variables that are reset to zero at the beginning of execution:
docseqno = 0 'Sequential number of Word docs in subdir
occur = 0    'Number of occurrences of searchterm in individual doc

'Writes the row heading for the word counts:
Cells(RowNoBottomList + 1, 1).FormulaR1C1 = "TotalWordsInDoc"
```

```
'Calling MS Word:
Set oWrd = CreateObject("Word.Application")
oWrd.Visible = False 'Saves time, avoids opening Word docs on screen
sNam = Dir(sPth & "*.doc") 'Read in all file names with extension .doc

'WHILE - WEND OUTERMOST LOOP GOING THROUGH ALL .DOC FILES IN SUBDIR:
    While sNam <> "" 'Loops until empty.
    oWrd.documents.Open sPth & sNam 'Opens one document at a time
    docseqno = docseqno + 1 'Keeps a sequential number on doc currently open

    'Copies file name to spreadsheet top row, first doc name to col 2, etc.:
    Cells(1, docseqno + 1).Value = sNam

    Dim rDcm As Object

        'RUN THROUGH SEARCH TERMS IN THE ACTIVE DOCUMENT
        For Counter1 = 2 To RowNoBottomList
        searchterm = Cells(Counter1, 1).Value

            'DO FIND OPERATION
            Set rDcm = oWrd.activedocument.Range
             '"Range" implies "search all" of this document.
            'The "Set .." statement has to be here, inside the Counter1
            'structure so that every time counter value updated, will search from
            'beginning of document:
            With rDcm.Find
            .Text = searchterm

                While .Execute
                rDcm.Select 'Selects another occurrence if there is one.
                occur = occur + 1 'Updates whenever a new occurrence found.
                Wend

            'Passes the occurrence count to the cell that is at the intersection
            'of search term (row) and document name (column) in the spreadsheet:
            Cells(Counter1, docseqno + 1).Value = occur
            occur = 0 'After passing the count,
                      'reset occurrence counter to zero for next search.

            End With
            ' END FIND OPERATION

        Next Counter1
        'COMPLETED RUNNING THROUGH ALL SEARCH TERMS WITHIN THIS DOCUMENT

        'WORD COUNT FOR EACH DOCUMENT
        'Places the total number of words (tokens, not types!) in the row
        'immediately below the word frequency matrix.
        'This statistic is needed for calculating relative word frequencies,
        'e.g., occurrences per 10,000 words.
        nowords = oWrd.activedocument.Words.Count
        Cells(RowNoBottomList + 1, docseqno + 1).Value = nowords
        'DONE WITH THIS DOCUMENT; IT CAN BE CLOSED:

    oWrd.activedocument.Close
    sNam = Dir

    Wend
    'END OUTERMOST LOOP. DONE WITH ALL WORD DOCS IN SUBDIR.

oWrd.Quit
Set oWrd = Nothing

End Sub
```

Sub FreqGivenTermsTextdocs()

```
'Excel macro to calculate frequency of search string in several .txt documents
'Search strings to use must be listed in leftmost column of active sheet,
```

```
'starting in row 2. (Cell R1C1 or A1 can be used as a title "SearchTerms")
'The macro writes the document names in the top row, starting in R1C2 (or "B2").
'In the row below the word frequency matrix, it writes the total word count
'for each document.

'Example of a segment of the resulting table:

'Searchstrings   text-001.txt text-002.txt text-003.txt
'Lutheran    1   9   3
'empower 1   0   0
'TotalWordsInDoc 374 659 683

'where "9" means that the word "Lutheran" occurs 9 times in document text-002.text,
'which is 659 words' long.


'Written by Aldo Benini, 5 June 2009
'Adapted from a similar macro for Word docs, FreqGivenTermsMSWorddocs(),
'which in part was adapted from a macro written by Ray Tweedale,
'"Import WILDCARD search to EXCEL from WORD", 2008

Dim sPth As String 'Path where .txt docs are stored
Dim sNam As String 'Name of .txt file
Dim oWrd As Object 'MS Word as an application object within Excel

'-----------------------

'PARAMETERS THAT YOU MUST SET:

'1. Subdirectory where .txt documents to investigate are stored:
sPth = "C:\Aldo\A_Automap\LWI2008\" 'Change as appropriate. Ensure backslash at the end.

'2. Column 1 (or "A" depending on cell reference type) holds the search terms
' from row 2 to the last row, without gap. Set the row number of the bottom cell
' (the one holding the last term in your list):
RowNoBottomList = 3 'Change as appropriate

'-----------------------
'Count variables that are reset to zero at the beginning of execution:
docseqno = 0 'Sequential number of Word docs in subdir
occur = 0    'Number of occurrences of searchterm in individual doc

'Writes the row heading for the word counts:
Cells(RowNoBottomList + 1, 1).FormulaR1C1 = "TotalWordsInDoc"

'Calling MS Word:
Set oWrd = CreateObject("Word.Application")
'We use MS Word to read and search these .txt documents
oWrd.Visible = False 'Saves time, avoids opening docs on screen
sNam = Dir(sPth & "*.txt") 'Read in all file names with extension .txt

'WHILE - WEND OUTERMOST LOOP GOING THROUGH ALL .txt FILES IN SUBDIR:
    While sNam <> "" 'Loops until empty.
    oWrd.documents.Open sPth & sNam 'Opens one document at a time
    docseqno = docseqno + 1 'Keeps a sequential number on doc currently open

    'Copies file name to spreadsheet top row, first doc name to col 2, etc.:
    Cells(1, docseqno + 1).Value = sNam

    Dim rDcm As Object

        'RUN THROUGH SEARCH TERMS IN THE ACTIVE DOCUMENT
        For Counter1 = 2 To RowNoBottomList
        searchterm = Cells(Counter1, 1).Value

            'DO FIND OPERATION
            Set rDcm = oWrd.activedocument.Range
             '"Range" implies "search all" of this document.
            'Has to be here, inside the Counter1
            'structure so that every time counter value updated, will search from
            'beginning of document:
```

```
            With rDcm.Find
            .Text = searchterm

                While .Execute
                rDcm.Select 'Selects another occurrence if there is one.
                occur = occur + 1 'Updates whenever a new occurrence found.
                Wend

            'Passes the occurrence count to the cell that is at the intersection
            'of search term (row) and document name (column) in the spreadsheet:
            Cells(Counter1, docseqno + 1).Value = occur
            occur = 0
           'After passing the count, reset occurrence counter to zero for next search.
            End With

            ' END FIND OPERATION

        Next Counter1
        'COMPLETED RUNNING THROUGH ALL SEARCH TERMS WITHIN THIS DOCUMENT

        'WORD COUNT FOR EACH DOCUMENT
        'Places the total number of words (tokens, not types!) in the row
        'immediately below the word frequency matrix.
        'This statistic is needed for calculating relative word frequencies,
        'e.g., occurrences per 10,000 words.
        nowords = oWrd.activedocument.Words.Count
        Cells(RowNoBottomList + 1, docseqno + 1).Value = nowords
        'DONE WITH THIS DOCUMENT; IT CAN BE CLOSED:

    oWrd.activedocument.Close
    sNam = Dir

    Wend
    'END OUTERMOST LOOP. DONE WITH ALL .txt DOCS IN SUBDIR.

oWrd.Quit
Set oWrd = Nothing

End Sub
```

Here ends the code of the two macros.

Association matrix

Explanations are given in comment lines of the code. The macro will not work unless a range is created "TermDocFreq" exactly as defined below.

Here starts the macro code:

```
Sub AssocMatrix()

'The macro produces an asymmetrical matrix of association coefficients between the terms
'used in a term - document frequency table: p(B | A) = (number of docs in which
'both A and B appear) / (number of docs in which A appears).

'Written by Aldo Benini 20 July 2009

'WHAT THE USER MUST DEFINE
'The user needs to name the range of the term - document frequency table "TermDocFreq".
'This range must hold the terms (in the leftmost column) and the doc names
'(in the top row) but must not include any cells that hold row (e.g. term occurrences
'in the corpus, on the far right)or column totals (e.g. word counts, in the bottom row).

Dim AssocCoef As Variant, DiagElem As Integer
```

```
'OPTIONAL PARAMETER TO SET DIAGONAL ELEMENTS TO ZERO
'The default produces diagonal elements p(A | A) = 1. For network visualization purposes,
'users may want to have these elements set to zero. The optional parameter

DiagElem = 1

'achieves that when it is set to zero.


'Inserts a new blank sheet to hold the association matrix
'and avoids using the same sheet name again if the user re-runs the macro:

nosheets = Sheets.Count
Set NewSheet = Sheets.Add(Type:=xlWorksheet)
NewSheet.Name = "Sheet" & (nosheets + 1) & "_AssocMatrix"
Sheets("Sheet" & (nosheets + 1) & "_AssocMatrix").Move after:=Sheets(nosheets + 1)

TDFrows = Range("TermDocFreq").Rows.Count
TDFcols = Range("TermDocFreq").Columns.Count

'Counters used in "For .. To" expressions:
'Counter1 goes through the terms of the TDF table to fetch the base term (A).
'Counter2 goes through the terms of the TDF table to fetch the target term (B).
'Counter3 goes through the documents to fetch the base and target term occurrences.
'Counter4 goes through the terms to find the diagonal elements of the association matrix
'if the user chooses to set them to zero.

For Counter1 = 2 To TDFrows

'Writes the row and column names of the association matrix (= copies the terms):
Cells(Counter1, 1).FormulaR1C1 = Range("TermDocFreq").Cells(Counter1, 1).FormulaR1C1
Cells(1, Counter1).FormulaR1C1 = Range("TermDocFreq").Cells(Counter1, 1).FormulaR1C1

    'Calculation of association coefficients starts here:
    For Counter2 = 2 To TDFrows

    'Resets these auxiliary variable each time Counter2 moves forward:
    AssocNumerat = 0
    AssocDenom = 0
    addproduct = 0
    mult1 = 0
    mult2 = 0

        For Counter3 = 2 To TDFcols

        If Range("TermDocFreq").Cells(Counter1, Counter3) > 0 Then
            mult1 = 1
        Else
            mult1 = 0
        End If

        If Range("TermDocFreq").Cells(Counter2, Counter3) > 0 Then
            mult2 = 1
        Else
            mult2 = 0
        End If

        addproduct = mult1 * mult2

        AssocNumerat = AssocNumerat + addproduct
        AssocDenom = AssocDenom + mult1
        Next Counter3

    'After Counter3 has come to its end, transfers the value to the matrix cell:
    AssocCoef = AssocNumerat / AssocDenom
    Cells(Counter1, Counter2) = AssocCoef
    'End calculation of coefficient for given values in Counter1, Counter2

    Next Counter2

'Next row in association matrix:
```

```
Next Counter1

'Formats coefficients uniformly:
    Range("A1").Select
    Range(Selection, ActiveCell.SpecialCells(xlLastCell)).Select
    Selection.NumberFormat = "0.00"
    Range("A1").Select

'Sets diagonal elements to zero if DiagElem = 0
    If DiagElem = 0 Then
        For Counter4 = 2 To TDFrows
        Cells(Counter4, Counter4) = 0
        Next Counter4
    Else
    End If

End Sub
```

Here ends the code for this macro.

Figure 8 on page 40, which uses an output of this macro, was drawn with the network visualization freeware "Pajek" (Batagelj and Mrvar 2009).

The Excel spreadsheet *"Benini_TextAnalysisMacros090721.xls"*, loaded with these macros, is available at www.aldo-benini.org .

References

Adolphs, S. (2006). Introducing electronic text analysis. New York, Routledge.

Altman, M. (2008). "The Impoverished Social Scientist's Guide to Free Statistical Software and Resources." Retrieved 5 June 2009, from http://maltman.hmdc.harvard.edu/micah_altman/socsci.shtml#TEXT.

Associated Press. (2010). "Arrests intensify Haiti adoption debate. Some groups call for moratorium, others fear long-term clampdown [1 February 2010]." Retrieved 1 March 2010, from http://www.msnbc.msn.com/id/35188317.

Bamberger, M., J. Rugh, et al. (2006). RealWorld evaluation : working under budget, time, data, and political constraints. Thousand Oaks, Sage Publications.

Batagelj, V. and A. Mrvar. (2009). "Pajek 1.24." Retrieved 19 July 2009, from http://pajek.imfm.si/doku.php.

Baulch, B. and L. Scott (2006). Report on CPRC Workshop on Panel Surveys and Life History Methods [Held at the Overseas Development Institute, London, 24-25th February 2006]. Manchester, Chronic Poverty Research Center.

Benini, A. (2008). Does Empowerment Work? Underlying concepts and the experience of two community empowerment programs in Cambodia and Tanzania. Washington DC.

Benoit, K., M. Laver, et al. (Undated). Wordscores for Stata and R, Wordscores.com.

Bryman, A. (2007). "Barriers to Integrating Quantitative and Qualitative Research." Journal of Mixed Methods Research **1**(1): 8-22.

Carley, K. M. (2009). AutoMap-2.7.70 [http://www.casos.cs.cmu.edu/projects/automap]. Pittsburgh, Carnegie Mellon University, Center for Computational Analysis of Social and Organizational Systems (CASOS).

Cortenraad, J. (2000). "Reflections from a refugee camp." Retrieved 4 May 2009, from http://www.unv.org/en/perspectives/doc/reflections-from-a-refugee.html.

Davies, R. (2005). "Scale, complexity and the representation of theories of change: Part II." Evaluation **11**(2): 133-149.

Denzin, N. K. and Y. S. Lincoln (2005). The Sage Handbook of Qualitative Research. Thousand Oaks, CA, Sage Publications Inc.

DiRT. (2009a). "Perform Qualitative Data Analysis." Retrieved 5 June 2009, from http://digitalresearchtools.pbworks.com/Perform+Qualitative+Data+Analysis.

DiRT. (2009b). "Text Analysis Tools." Retrieved 5 June 2009, from http://digitalresearchtools.pbworks.com/Text+Analysis+Tools.

Duriau, V. J., R. K. Reger, et al. (2007). "A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements." Organizational Research Methods **10**(1): 5.

Feinerer, I., K. Hornik, et al. (2008). "Text mining infrastructure in R [http://www.jstatsoft.org/v25/i05/paper]." Journal of Statistical Software **25**(5): 1-54.

Feinerer, I. and F. Wild. (2009). "CRAN Task View: Natural Language Processing." Retrieved 7 June 2009, from http://cran.r-project.org/web/views/NaturalLanguageProcessing.html.

Goldman, M. (2001). "The Birth of a Discipline: Producing Authoritative Green Knowledge, World Bank-Style." Ethnography **2**(2): 191-217.
Haynes, R. (2009). "Preparing Text Data for Statistical Analysis using Excel." Retrieved 5 June 2009, from A Lean Six Sigma Tools Blog For Lean Six Sigma Professionals.
Hüning, M. (2010, 28 December 2009). "TextSTAT 2.8g for Windows." Retrieved 23 February 2010, from http://www.niederlandistik.fu-berlin.de/textstat/software-en.html.
Hwang, S. (2008). "Utilizing Qualitative Data Analysis Software: A Review of Atlas. ti." Social Science Computer Review **26**(4): 519-527.
Landmann, J. and C. Zuell (2008). "Identifying Events Using Computer-Assisted Text Analysis." Social Science Computer Review **26**(4): 483-497.
LCL. (2009). "TermExtractor [Demo version]." Retrieved 11 June 2009, from http://lcl.uniroma1.it/termextractor/.
Leech, G., P. Rayson, et al. (2001a). "Companion Website: Frequency lists [List 1.1.: Alphabetical frequency list of the whole corpus (lemmatized) and key]." Retrieved 20 April 2009, from http://ucrel.lancs.ac.uk/bncfreq/flists.html.
Leech, G., P. Rayson, et al. (2001b). Word frequencies in written and spoken English: based on the British National Corpus. London, Longman.
Leser, U. (2008). "Text Analytics: Evaluating IR Systems and Text Preprocessing [Lecture notes]." from www.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/sose08/hk_text/03_evaluation_normalization.pdf.
Lewis, D. (1997). "Reuters-21578 Text Categorization Collection Distribution 1.0." Retrieved 7 June 2009, from kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.
Lewis, R. B. and S. M. Maas (2007). "QDA Miner 2.0: Mixed-Model Qualitative Data Analysis Software." Field Methods **19**(1): 87-108.
Lowe, W. (2008). "Understanding Wordscores." Political Analysis **16**(4): 356-371.
Luhmann, N. (1987). Soziale Systeme. Grundriss einer allgemeinen Theorie. Frankfurt am Main, Suhrkamp.
Luhmann, N. (1997). Die Gesellschaft der Gesellschaft. Frankfurt am Main, Suhrkamp.
Lurie, N. H. (2004). "Decision making in information-rich environments: The role of information structure." Journal of Consumer Research **30**(4): 473-486.
McRitchie, D. (2008). "My Excel Pages." Retrieved 1 March 2010, from http://www.mvps.org/dmcritchie/excel/excel.htm
Moore, S., E. Eng, et al. (2003). "International NGOs and the role of network centrality in humanitarian aid operations: A case study of coordination during the 2000 Mozambique floods." Disasters **27**(4): 305-318.
Oxfam. (2008). "Annual Report 2007." Retrieved 10 June 2009, from http://www.oxfam.org/sites/www.oxfam.org/files/OI-annual-report-2007-en.pdf.
Popping, R. (2000). Computer-assisted text analysis. London ; Thousand Oaks, Calif., Sage Publications.
Rockwell, G. (2003). "What is text analysis, really?" Literary and Linguistic Computing **18**(2): 209-219.

Sclano, F. and P. Velardi (2007). Termextractor: a web application to learn the shared terminology of emergent web communities. Rome, University of Roma “La Sapienza”.
Smith, M. and et.al. (2010). NodeXL. Network Overview Discovery and Exploration for Excel 2007, CodePlex Open Source Community.
StataCorp (2007). STATA Statistical Software: Release 10. College Station, TX, StataCorp LP.
The Economist (2010). "Data, data everywhere. A special report on managing information." (February 27th, 2010).
van Atteveldt, W. (2008). Semantic Network Analysis. Techniques for Extracting, Representing and Querying Media Content. Charleston SC, BookSurge Publishing.

About the author

Aldo Benini has a dual career in rural development, with a focus on Bangladesh and another on organizations of the poor, and in humanitarian action. In the latter capacity, he has worked for the International Committee of the Red Cross and for the Global Landmine Survey. He has a Ph.D. in sociology from the University of Bielefeld, Germany, based on field research in community development in West Africa.

Benini is a citizen of Switzerland and an independent researcher based in Washington DC. He can be contacted at aldobenini [at] gmail.com. Publications are available at http://aldo-benini.org .