BOOK APPENDIX I: METHODLOGY

This methodology appendix contains two parts. The first provides a detailed explanation of the Alceste textual analysis software—its strengths, its weaknesses and its merits relative to other such software. (Some of this first section derives from (Schonhardt-Bailey, Yager et al. 2012). The second part sets out our approach to preparing the various corpora for purposes of analysis. That is, we explain our editing of FOMC transcripts and congressional hearings, particularly the means by which we have sought to control the lemmatization process employed by the Alceste software.

I. Alceste

a. Overview

Alceste was developed by Max Reinert (Reinert 1983; Reinert 1998; Reinert 2003) and has been applied in sociology, psychology, and political science (Noel-Jorand, Reinert et al. 1995; Lahlou 1996; Noel-Jorand, Reinert et al. 1997; Brugidou 1998; Guerin-Pace 1998; Bauer 2000; Brugidou 2003; Noel-Jorand, Reinert et al. 2004; Schonhardt-Bailey 2005; Schonhardt-Bailey 2006; Bara, Weale et al. 2007). Its specific advantages have been discussed elsewhere (Jenny 1997; Brugidou, Escoffier et al. 2000). (See also http://www.cmh.pro.ens.fr/bms/arcati/BMS54-Jenny-New.htm for a list of the websites and papers comparing text mining software.) An open-source reproduction of Alceste is available in the Iramuteq software (http://www.iramuteq.org/).

Alceste was initially designed to measure what Max Reinert calls the “lexical worlds”. As Reinert explains: “… we assume that the speaker, during his speech, is investing successive different worlds and these worlds, by imposing their properties, thereby impose a specific vocabulary. Therefore, the statistical study of the distribution of this vocabulary should be able to trace these ‘mental rooms’ that the speaker has successively inhabited; traces perceptible in terms of ‘lexical worlds’…" (Reinert 1987). In other words, a lexical world is a specific vocabulary, which inherits its properties from what the subject is talking about—e.g., if the text is about medicine, there will be many medical terms. Conversely, if there are many medical terms, this is a cue that the text may be about medicine.

By purely distributional means the sets of words that go together in the discourse are isolated and represented to the researcher, as a trace of some “lexical world” which remains to be interpreted. The software accomplishes this using only a statistical approach to analyze the distribution of words in the corpus, while remaining completely deaf to the meaning of words themselves. The only semantic aspects inbuilt in the software are some grammatical dictionaries which enable reducing verbal forms to a single root (reducing all the flexions of a single verb to its radical, or names in plural to singular), and classifying words into various grammatical classes (nouns, verbs, articles etc.) so as to eliminate function words (articles, some prepositions) in the analysis.

The basic idea of the software is to find “lexical worlds” in the speaker’s discourse. As a metaphor, consider a tourist who visits a country where there is a seaside, a town, a desert and a forest. The tourist stays in this country for a month and goes to various places; maybe several times a day to the town, every second day to the seaside and the rest of the time in the forest, but very few times in the desert. Every 10 minutes, the tourist posts what he sees around him on his internet blog. Let us organize his posts by putting together those which share the same lexical content (e.g., the ones containing “building”, “street” in one class; the ones containing “trees”, “plants” in another; the ones with “sea”, “beach” in a third, etc.). In doing so, even without knowing the country we will reconstruct through these lexical associations classes corresponding to the areas to which they refer (town, forest, seaside, desert) and understand what these areas are based on the lexical content of the classes (the class containing “sea” and “beach” probably refers to the seaside, etc.). It is because these words co-occur locally that we can make interpretations (a tree alone means nothing, and could be found in towns; but a tree co-occurring with other trees and with plants means forest). We shall also be able to assess in which of these areas the tourist stayed more often. And if we did that with two tourists, we may find some differences in the places where they tend to stay (assuming that if one was more often at the seaside and the other in the forest, this would be reflected in their discourse).

In the same way, as politicians or policymakers run through semantic fields when producing discourse, their statements provide us with a lexical distribution that reflects the content of these fields (their lexical worlds). By classifying together the statements that contain similar words, we can hope to understand what semantic territories were behind the construction of the observed discourse.

Alceste operationalizes these notions of “statements”, “words”, and “similarity”. Statements are approximated by “Elementary Context Units” (ECUs), which are natural sentences or natural fragments of sentences delimited by punctuation so as to have similar length. Alceste constructs a dictionary of “lexical forms” (“lexemes”) which are lemmatized words, more useful for our purpose in terms of semantics.

To assess similarity between statements, Alceste constructs a matrix that crosses ECUs and lexemes, where the cells sign the presence or absence of that lexeme in the ECU. Alceste then operates on this matrix a descending classification, which produces classes of similar context units. The descending classification technique used maximizes the similarity between statements in the same class *and* also maximizes the difference between the classes.

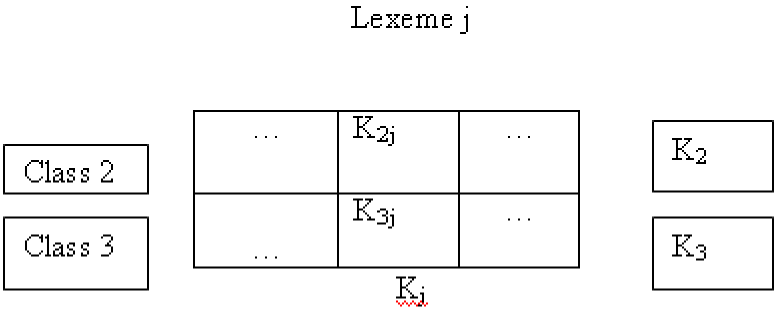
In the end, the analyst is provided with a series of classes and of statistical cues in the form of typical words, typical ECUs, typical authors and so on. This provides basis for "interpreting" the classes as lexical worlds.

Operations in Alceste are statistical, transparent and reproducible, until the final moment of interpretation, where the analyst assigns a label to each set of specific vocabulary which was identified as a lexical world by the software, on the basis of co-occurrences and distribution patterns.

b. The Principle of Descending Classification

The core of Alceste is based on descending classification of text segments. The objective is to sort the ECUs in a partition of classes which are each homogeneous and as different as possible from one another. The initial table consists of as many lines as ECUs, and as many columns as lexemes chosen for analysis. At the intersection of a row i and column j the value is either 1 if ECU_i contains at least one occurrence of the lexeme_j ; 0 otherwise.

The classification is a recursive algorithm. The first class comprises the total set of context units. The program then attempts to partition that class into two further classes which are each as homogeneous as possible and as different as possible from one another. This overlap between classes can be measured by the χ^2 of a table with 2 rows (one for each class) and as many columns as there are lexemes. A cell at the intersection of row i and column j will contain the number k_{ij} of context units of class I containing lexeme j.



With, for example:

$$k_2 j = \sum_{i \in I_2} k_{ij}; k_2 = \sum_{i \in I_1} k_{2j}; k_j = k_{2j} + k_{3j};$$

The χ^2 can written as:

$$chi2 = k_2.k_3 \sum_{j \in J} (k_{2j} / k_2 - k_{3j} / k_3)^2 / k_j$$

The objective is to search, among all possible partitions in two classes, for the one that maximizes the χ^2 of this table. In practice, the algorithm

a) Calculates the first factor of the correspondence factor analysis of the table (R_j space, with the χ^2 metrics (Benzecri 1973);

b) Slides the orthogonal hyperplane along that axis until it reaches a position that maximizes the inter-class inertia of the two parts of the cloud of context units that are separated by this hyperplane;

c) Optimizes the partition with a local exchange algorithm by swapping context units individually around the hyperplane.

Following an iterative process, the descending hierarchical classification method decomposes the classes until a predetermined number of iterations fails to result in further significant divisions. The result is a hierarchy of classes, which may be schematized as a tree diagram.

One advantage in using of descending classification (e.g., over ascending clustering techniques) is robustness. Textual data matrices are “scarce matrices” (mostly constituted of zeros, since only a small part of the whole vocabulary is used in each sentence) and as such, are very sensitive to artifacts. A few very similar sentences (e.g., using a fixed expression) will create a strong correlation that can “pull” an artifactual class. In the descending classification algorithm used in Alceste, such effects stay local and do not propagate to the whole analysis.

c. The Four Steps of the Method

Alceste has been described as a "methodology" insofar as it "integrates a multitude of highly sophisticated statistical methods," (Kronberger and Wagner 2000: 306). The details of the algorithms has been extensively described (Reinert 1983; Reinert 1987; Reinert 1990; Beaudouin and Lahlou 1993; Reinert 1993; Lebart and Salem 1994; Lahlou 1995a; Reinert 1998; Reinert 2003). Below is a short summary.

Step A : The dictionaries of the vocabulary are created. They identify word categories (verbs, nouns, etc.) which are useful because only some categories are to be used in analysis (excluding "tool words" such as articles or pronouns). Vocabulary is parsed using a grammatical dictionary, and reduced into a dictionary of (lemmatized) lexical forms. All forms of a verb are converted to their infinitive, plurals are reduced to singular, and some variants of the same lexical root are reduced to the root. These forms are the basic "lexemes" upon which calculations are made.

Couples (two consecutive forms) and repeated segments (consecutive series of forms) are also found and listed in this step. They are useful in illustrating the classes at a later step. Calculations of frequencies of the forms are also done at this stage, which may be used in further steps, e.g., for eliminating from the analysis words which occur only once.

Step B: The software cuts the text into "statements". These are obtained empirically as Elementary Context Units ("ECUs"), in practice sentences or parts of sentences cut by natural punctuation. There are three types of context units. Initial Context Units or "ICUs" are the pieces of text that constitute the corpus to be analyzed (here, the speeches or statements of politicians and policymakers). The ICU is essentially the sampling unit—i.e., a pre-existing division of the text and is specified by the user.

Elementary Context Units ("ECU") are the atomic pieces of text. The ECU is a "gauged sentence" (or recording unit), which the program automatically constructs based upon and punctuation in the text. An ECU is delimited by a punctuation sign, and contains at least 15 occurrences of words.

In the course of segmentation, the software will also use somewhat larger Context Units (CUs), made by concatenation of several succeeding ECUs.

In text analysis, a persistent and difficult issue concerns the optimal length of a "statement" as a semantic unit. Various possible solutions for a contextual unit would be a sentence, a paragraph, a piece of sentence, and so on. Alceste resolves this issue by not trying to identify directly the statement length; rather it produces classifications that are *independent* of the length of the statements.

For this Alceste creates two classifications, each using units of different lengths of CUs, and then retains only the classes which appear in both classifications: these classes are independent of the length of statements. In practice, CUs in each classification are constructed as concatenations of ECUs which have a minimal length of N1 *active words* (e.g., 12), and in the second classification a length of N2 active words (e.g., 18). An active word is a lexical root or “lexeme” (see Step A) of a noun, verb or adverb which occurs at least two times in the corpus. Alceste compares the two classifications, on the basis of ECUs in the classes. Only the classes that appear in both classifications are retained for analysis. The resulting classification is stable, and, as said, independent of the length of the CUs. This leaves a number of ECUs unclassified; thereby approximating a measure of goodness-of-fit.

The stability of the partitioning is measured by constructing a table that crosses all the classes (including all levels of nodes) obtained in the first classification and all the classes obtained in the second classification. In each cell (Cpq) of this table is the χ^2 associating the two classes p and q. The result is a “signed chi-square table”—that is, a data table with the positive and negative links between the classes. This table helps to select the classes which share the higher number of ECUs. Interestingly, this table not only enables the program to retain the stable partitions, but also to obtain an empirical solution as where to stop descending classification in the classification tree: when stems are not stable, the tree truncates at the higher node.

This description, although limited, provides an indication of the elegance of the algorithm and the statistical underpinnings necessary to cope with the specific type of matrices encountered in textual data analysis (“scarce matrices”). This accounts for this specific software’s exceptional robustness. For a detailed exposition of the algorithm, see (Reinert 1983; Lahlou 1995a; Bastin 2002); for a step-by-step explanation of each stage of the analysis, see (Reinert 1990), and for a simple illustration (Kronberger and Wagner 2000).

Steps C and D provide various auxiliary calculations to assist in the interpretation and description of classes. Most suggestive are the lists of the most representative vocabulary of each class (lexemes ordered by decreasing χ^2), and selected ECUs that best represent the class, based on their lexical content.

A correspondence analysis is also performed in this step to provide a comprehensive representation of the semantic field that situates the relative positions of the classes and lexemes.

d. Alceste: Strengths and Weaknesses

The use of text-mining or text analysis software has proliferated in recent years, both in the popular and academic literatures. A survey of these software lies outside the scope of this paper; here, we confine our discussion to computer-assisted content analysis.

One form of automated content analysis is topic modelling, where the task is automatically to classify the contents of documents into "topics". Each topic is understood to comprise a probability distribution over a fixed vocabulary of words or terms, and each document exhibits any number of topics in different proportions (Blei and Lafferty 2009). The basic idea is that words are indicative of topical content, and the task is to map the words into topics using a specified parametric form. These models are useful for exploring and cataloguing vast digital libraries (Blei and Lafferty 2006; Blei and Lafferty 2009), or for categorizing a large number of speeches on a variety of subjects, where very little substantive knowledge of the subjects themselves is required (Quinn, Monroe et al. 2010). A clear disadvantage of this form of content analysis is that it effectively discards a large amount of textual information. To illustrate, using this approach for corpora such as those we have analysed in this book, each speech of politician would have been assigned to a topic (and just one topic) based on the largest proportion of words in the speech. Let's say that a given speech contains four components (topics), each with shares of 40%, 30%, 20% and 10% (and assume that 100% of the speech is successfully classified into topics). The whole of the speech would be allocated to the first topic, given that its share is the largest. Hence, 60% of the speech would be misallocated. In short, each speech is allocated to just one topic.

A second approach to content analysis assumes that speakers or authors of textual data convey meaning in a more thematic fashion, and so it is not just the words that help to classify content, but also the context in which the words appear. Rather than conceptualizing words in a univariate distributional pattern (e.g., as in topic modelling), a thematic approach examines the bivariate associations between words and sets of words in order to map out so-called "lexical worlds", and the relationships between lexical worlds within a single corpus. As discussed in the first part of this appendix, Alceste uses this thematic approach. A key feature of Alceste is that it can be used to identify the speakers' tendency to articulate particular ideas and arguments—ideas and arguments which can then be correlated with characteristics of the speaker (e.g., in Chapters 3 and 4, these characteristics have included the name the name of speaker, party affiliation, the role of the committee member and so on).

Alceste does not require any pre-coding but its application is constrained in that it cannot analyze very large corpora. (Although subsequent versions may allow a larger corpus, Alceste 4.7 requires that the corpus not exceed 15 mb.) In contrast to applications of topic modeling where, for instance, vast libraries of documents with diverse topics might be analyzed, Alceste is suited to more focused research projects—for example, documents relating to a particular area of policy (monetary policy, trade policy, health), or the speeches of politicians which may be expected to contain broadly similar themes (State of Union addresses). Moreover, whereas topic modeling requires very little substantive knowledge of the topics themselves, Alceste is most effective when it is joined with expert substantive knowledge of the subject matter, since contextual knowledge is often essential for interpreting the form of argumentation as well for extending the analysis into its more specialist usages (e.g., the Cross Data Analysis, or *Tri-Croisé*), as seen in our Chapters 3 and 4.

Although Alceste is useful and widely used by practitioners and researchers, the practical implementation of the method is not readily transparent to the users. One expert statistician in textual analysis (Ludovic Lebart) has commented in email exchanges to the authors (June 2011) that it is "an astute and interesting approach to text analysis, with an undeniable heuristic value … which allows for immediate and unprejudiced explorations of the corpus." But he also notes that there are also a series of decisions (default values, frequency thresholds, selection of statistical criteria, clustering methodology) for which the proofs are not readily apparent. While the settings of the software are "devised to ensure a smooth functioning of the algorithm, to obtain a balanced set of clusters, to present legible graphical displays, [and] the resulting product is easy to use [this] is not a guarantee of quality from a statistical point of view." For this reason, we conduct two extensive robustness tests. The first is the series of interviews discussed in our Chapter 5, while the second is a re-analysis of the primary corpora using another textual analysis software which is similar but still distinct from Alceste. We present this in our Book Appendix IV.

I. Preparation and Editing of Our Corpora

Alceste requires a minimal amount of editing of the text prior to analysis, which is detailed elsewhere (Reinert 1998). However, our FOMC and congressional hearing transcripts required a finer specification in order to assure stable results and to avoid errors in the lemmatization process.

First, in Alceste, a capital letter followed by a lower case letter is automatically changed to a lower case letter; however, a word constituted by capital letters only (as an acronym) remains unchanged. So, for instance, the word “Fed” (as in Federal Reserve) would be changed to “fed” and thus read by the software as the past tense of “feed”. (Aside from the obvious distortion of its meaning, the word would be treated as a verb rather than a noun.) To avoid this and other potential distortions in the lemmatization process, we had to make the necessary substitutions prior to analysis (e.g., “Fed” would become “Federal Reserve”).

Second, a hyphen is not recognised as a liaison link by the software, so for instance “Y-2-K” would be read as the separate letters “Y 2 K” rather than as a single phrase. In these cases, the hyphen is replaced with an underscore (“Y_2_K”).

Hence, as a means to avoid distortions from the lemmatization process, we edited the FOMC and congressional hearing transcripts as follows:

- All names are joined with hyphens (“Arthur_Burns”).
- Countries, regions, states are similarly joined, as needed (Southeast_Asia, United_Kingdom).
- Key institutions and phrases are changed as follows:

New York Federal Reserve (New_York_Federal_Reserve)
Fed or Federal Reserve (Federal_Reserve)
Bank of Japan (Bank_of_Japan)
Domestic Open Market Operations (Domestic_Open_Market_Operations)
Fannie Mae (changed to Fannie_Mae)
Freddie Mac (Freddie_Mac)
G5 (G_5)
G7 (G_7)
Capital investment (capital_investment)
M1 (M_1)
M2 (M_2)
M3 (M_3)
Home Loan Bank/s (Home_Loan_Banks/s)
Federal Open Market Committee/FOMC (F_O_M_C)
Consumer Price Index/CPI (C_P_I)
European Central Bank/ECB (E_C_B)
Y-2-K (Y_2_K)
GDP (G_D_P)
Latin America/Latin American (Latin_America/n)
UK/United Kingdom/Britain (United_Kingdom)
Bank of England (Bank_of_England)
Brazilian Central Bank (Brazilian_Central_Bank)

NASDAQ (N_A_S_D_A_Q)
European Union (European_Union)
Exchange rate/s (exchange_rate; or exchange_rates)
Interest rate/s (interest_rate; interest_rates)
Inflation/inflation rate (inflation_rate; inflation_rates)
Monetary policy (monetary_policy)
Unemployment/unemployment rate (unemployment_rate; unemployment_rates)
Labor market (labor_market)
Manufacturing sector (manufacturing_sector)
Bond market (bond_market)
NAIRU (N_A_I_R_U)
Federal funds rate (Fed_Funds_Rate)
Federal Reserve Act (Federal_Reserve_Act)
Oil prices (oil_prices)
Philips curve (Philips_curve)
Taylor rule (Taylor_rule)
Humphrey-Hawkins (Humphrey_Hawkins)
Salomon Brothers (Salomon_Brothers)
Foreign Currency Operations (Foreign_Currency_Operations)
National economy (national_economy)
Financial markets (financial_markets)
Hong Kong Shanghai Bank (Hong_Kong_Shanghai_Bank)
Senate Banking Committee (Senate_Banking_Committee)
House Banking Committee (House_Banking_Committee)
System Open Market Account (System_Open Market_Account)

(1980s congressional hearings)

Monetary targets (monetary_targets)
Budget deficit/s (budget_deficit)
Gramm-Rudman-Hollings (Gramm_Rudman_Hollings)
Baker plan (Baker_plan)
Federal deficit/s (federal_deficit)
Federal spending (federal_spending)

(2006-08 files)

sub-prime mortgage (sub_prime_mortgate)
sub-prime lending (sub_prime_lending)
sub-prime market (sub_prime_market)
credit crunch (credit_crunch)
Northern Rock (Northern_Rock)
Sarbanes-Oxley (Sarbanes_Oxley)

Beyond these more technical alterations to the text, we also make minor substantive changes which, in our view, allow a "cleaner" analysis of the text. Our intent here is to delete superfluous and (for our purposes) irrelevant text. These are more prevalent for the FOMC meetings than for the congressional hearings. For example, given our focus on monetary policy in the FOMC, unwanted text relates to procedural and formal dialogue (e.g., administration of the meeting, including lunch, coffee, the order of speaking); formal introductions and so on, largely by the chairman; jokes and general banter; sporting metaphors unless they really do illuminate policy; remarks of no substance back to the chairman (including flattery of him); ratification of the minutes of the previous meeting; the annual confirmation of appointments of chair, vice chair, system manager; discussion of detailed intervention operations by the New York Fed as well as of data which adds next to nothing ("Is that number measuring Wednesday to Wednesday or something else"); and discussion of issues unrelated to the monetary policy decision (e.g., release of FOMC papers to the public).

Bara, J., A. Weale, et al. (2007). "Analysing Parliamentary Debate with Computer Assistance." Swiss Political Science Review **13**(4): 577-605.

Bastin, G. (2002). Note sur la méthode Alceste. Mondes sociaux et mondes lexicaux.(Note on the Alceste method. Social worlds and lexical worlds), Melissa, http://www.melissa.ens-cachan.fr/spip.php?article200. (accessed April 13, 2012).

Bauer, M. (2000). Classical Content Analysis: A Review. Qualitative Researching with Text, Image and Sound: A Practical Handbook. M. W. Bauer and G. Gaskell. London, Sage Publications**:** 131-151.

Beaudouin, V. and S. Lahlou (1993). L'analyse lexicale, outil d'exploration des représentations. Réflexions illustrées par une quinzaine d'analyses de corpus d'origines très diverses (Lexical analysis, a tool for exploring representations. Reflections illustrated by fifteen analyses of very diverse corpuses). Cahiers de Recherche du Crédoc. Paris, Crédoc. **n°48, septembre** 146

Benzecri, J.-P. (1973). L'analyse des données. Tome1: La Taxinomie. Tome 2: L'Analyse des Correspondances (Data analysis. Volume 1 : Taxonomy. Volume 2 : Correspondance Analysis). Paris, Dunod.

Blei, D. M. and J. D. Lafferty (2006). Dynamic Topic Models. 23rd International Conference on Machine Learning. Pittsburgh, PA.

Blei, D. M. and J. D. Lafferty (2009). Topic Models. Text Mining: Classification, Clustering, and Applications. A. Srivastava and M. Sahami. Boca Raton, FL, CRC Press**:** 71-94.

Brugidou, M. (1998). ""Epitaphes, l'image de Francois Mitterrand à travers l'analyse d'une question ouverte posée à sa mort (Epitaphs, Francois Mitterrand's Image: An Analysis of an Open Question Asked on His Death)." Revue Française de Science Politique **48**(1): 97-120.

Brugidou, M. (2003). "Argumentation and Values: An Analysis of Ordinary Political Competence Via An Open-Ended Question." International Journal of Public Opinion Research **15**(4): 413-430.

Brugidou, M., C. Escoffier, et al. (2000). "Les facteurs de choix et d'utilisation de logiciels d'analyse de données textuelles (parameters for chosing and using text mining software)." Journées internationales d'analyse statistiques des données textuelles http://lexicometrica.univ-paris3.fr/jadt/jadt2000/tocJADT2000.htm (accessed April 18, 2012) **JADT 2000**: 373-380.

Guerin-Pace, F. (1998). "Textual Analysis, An Exploratory Tool for the Social Sciences." Population: An English Selection, special issue of New Methodological Approaches in the Social Sciences **10**(1): 73-95.

Jenny, J. (1997). "Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification. (Methods and formalized practices for content and discourse analysis in contemporary French sociological research. State of the art and proposal for a taxonomy)." Bulletin de Méthodologie Sociologique (B.M.S.) **54**(March): 64-112.

Kronberger, N. and W. Wagner (2000). Keywords in Context: Statistical Analysis of Text Features. Qualitative Researching with Text, Image and Sound: A Practical Handbook. M. W. Bauer and G. Gaskell. London, Sage Publications**:** 299-317.

Lahlou, S. (1995a). Penser Manger. Les représentations sociales de l'alimentation (Food for Thought : Social Representations of Eating). Paris: EHESS, École des Hautes Études en Sciences Sociales. **Ph.D. diss.:** 735.

Lahlou, S. (1996). "A method to extract social representations from linguistic corpora." Japanese Journal of Experimental Social Psychology **35**(3): 278-291.

Lebart, L. and A. Salem (1994). Statistique Textuelle (Statistics for Text Mining). Paris, Dunod.

Noel-Jorand, M. C., M. Reinert, et al. (1995). "Discourse analysis and psychological adaptation to high altitude hypoxia." Stress Medicine **11**: 27-39.

Noel-Jorand, M. C., M. Reinert, et al. (1997). "A New Approach to Discourse Analysis in Psychiatry, applied to Schizophrenic Patient Speech." Schizophrenia Research **25**: 183-198.

Noel-Jorand, M. C., M. Reinert, et al. (2004). "Schizophrenia: The Quest for a Minimum Sense of Identity to Ward Off Delusional Psychosis." The Canadian Journal of Psychiatry **49**(6): 394-398.

Quinn, K. M., B. L. Monroe, et al. (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs." AMerican Journal of Political Science **54**(1): 209-228.

Reinert, M. (1983). "Une methode de classification descendante hierarchique: application a l'analyse lexicale par contexte." Les Cahiers de l'Analyse des Donnees **8**(2): 187-198.

Reinert, M. (1987). "Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud (Descending Hierarchical Classification and context-based lexical analysis : Application to the corpus of poems by A. Rimbaud)." Bulletin de Méthodologie Sociologique **13**(janvier): 53-90.

Reinert, M. (1990). "ALCESTE. Une methodologie d'analyse des donnees textuelles et une application: Aurelia de Gerard de Nerval." Bulletin de Methodologie Sociologique **26**: 24-54.

Reinert, M. (1993). "Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. ("Lexical Worlds" and their "logics" through the statistical analysis of a corpus of narratives of nightmares)." Langage et Société **66**: 5-39.

Reinert, M. (1998). ALCESTE users' manuel (English version). Toulouse, Image.

Reinert, M. (1998). Quel objet pour une analyse statistique du discours ? Quelques réflexions à propos de la réponse Alceste (What is the object of a statistical analysis of discourse? Some reflections about the Alceste solution). Proceedings of the 4th JADT (Journées d'Analyse des Données Textuelles) Université de Nice JADT.

Reinert, M. (2003). "Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste (The function of repetition in the representation of meaning and its statistical approach in the Alceste method)." Semiotica **147**(1-4): 389-420.

Schonhardt-Bailey, C. (2005). "Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches." PS: Political Science and Politics **38**(3): 701-711.

Schonhardt-Bailey, C. (2006). From the Corn Laws to Free Trade: Interests, Ideas and Institutions in Historical Perspective. Cambridge, MA, MIT Press.

Schonhardt-Bailey, C., E. Yager, et al. (2012). "Yes, Ronald Reagan's Rhetoric was Unique—But Statistically, How Unique?" Presidential Studies Quarterly(September).