

Standardised category systems in computer aided content analysis

An overview

Dr. Harald Klein, M.A.
Social Science Consulting
Lutherstr. 2
49082 Osnabrück
Germany

info@social-science-consulting.com

1 Abstract

One of the most time consuming parts in a content analysis is the construction of a category system. Most researchers have to develop their own category system, and often this will never be used again. This paper will show that standardised category system for general and specific research purposes are available. The category systems and their categories will be discussed, and one focus is which category systems can be used with which software.

Although many content analyses using the dictionary approach are performed, the category system is often used once only. So a lot of work is wasted, researchers may be forced to re-invent the wheel. This paper's purpose is to show that there are standardised category system available and encourages researchers to publish their category systems and make these available for further use.

The <http://www.textanalysis.info> web site gives more information on this topic and the related software for text analyses. This paper was presented at the RC33 of the ISA in Amsterdam, 2004.

2 Introduction

This paper gives an overview of already existing category systems. The description contains the name of the category system, the year of publication, and the categories. Also the software programs that can be used with the category system is included. The purpose of this paper is to document the category systems and how to make use of them. Although some of them are in digitised files, that does not mean every text analysis programs can use it.

A computer aided content analysis requires a category system. It contains categories, and each category consists of search patterns. Each search pattern must be a valid indicator for the category it represents. The construction of the category system and finding valid search patterns is a very time consuming work. Therefore it is desirable to use already existing category systems.

Some aspects are important for the decision what category systems might be used:

- Is the category system general or specific? Or in other words: can the category system be used for different text genres?
- Can the category system be edited by the researcher, thus allowing the porting to other programs?
- Is the category system available for different languages? If this is the case, international comparisons are possible. It also means that translation problems must be solved, too.
- Does disambiguation of ambiguous search patterns take place? Many search patterns are ambiguous, their meaning depends on the context. Most programs cannot differentiate between meanings, but if disambiguation is used, the validity of the category systems rises.
- What software can be used to apply the category system? If the category systems are editable, they can be used with other software. Often the form of the category systems requires intensive editing or even typing the search patterns again. Sometimes small converting programs exist that convert a category file into another format (e.g. from Textpack to TextQuest).

Quite a few category systems exist, but some are not publicly available nor published, and therefore these category systems are not listed here. The category system are either built-in and come with the software, or they can be used with more than one program. Currently I am aware of the following category systems, this list may be incomplete:

- Harvard: Harvard IV dictionary by Phil Stone et al.
- Lasswell value dictionary: by J. Zvi Namenwirth and Robert Philip Weber
- RID: Regressive Imagery Dictionary by Colin Martindale
- MCCA: Minnesota Contextual content analysis by Don McTavish and Ellen B. Pirro
- LIWC: Linguistic word count by James W. Pennebaker
- PCAD by Louis Gottschalk
- DAL: Dictionary of Affect in Language by Cynthia M. Whissell
- Forest value: by David N. Bengston and Zhi Xu
- verbal tone of political messages by Roderick F. Hart
- HKW: Hamburger kommunikationssoziologisches Wörterbuch (Hamburg's communication-sociological dictionary) by Alexander Deichsel
- DAW: Dresdner Angstwörterbuch (Dresden anxiety scale) by Hendrik Berth
- Umweltschutzwörterbuch (environment): by Klaus Schönbach, Elmar Bröker, and Volker Schlöggell
- news factors: by Harald Klein
- personal ads: by Helmut Giegler and Harald Klein

The following table gives an overview on the features of the category system. Std means, that the results of the coding are compared with built-in standards of the program. Disamb means, that the category systems works with disambiguation rules or not. If it does, it can identify the correct meaning of a search pattern, but is limited to one language (English).

Figure 1: Features of category systems

category system	general	editable	std	disam	languages
Harvard IV	yes	yes	no	yes	English, French
Lasswell value	yes	yes	no	yes	English
RID	yes	yes	no	no	English, French, Portuguese, German, Swedish, Latin
MCCA	yes	yes	yes	yes	English
LIWC	yes	yes	yes	no	English, Italian, Spanish, Dutch, German, Korean
verbal tone	no	no	yes	yes	English
anxiety scales	no	no	no	no	English
DAL	yes	yes	yes	no	English. French, Portugese
Forest value	no	yes	no	no	English
DAW	no	yes	no	no	German
HKW	yes	yes	no	no	German
ecological	no	yes	no	no	German
personal ads	no	yes	no	no	German
news factors	no	yes	no	no	German

After the decisions due to technical aspects, one has to choose what kind of category system is useful. There are two types of category systems:

- general: these can be used for all kind of texts and allow the re-analysis of texts.
- specific: they are developed for one kind of text only. Often these category system are built-in into software so it is difficult to purpose only.

3 The dictionaries

3.1 Harvard IV Psychological dictionary

The Harvard IV Psychological dictionary is one of the dictionaries that work with the General Inquirer/Mew Inquirer. It was ported to Protan and TextQuest also. The category system consists of 83 unique categories, the rest of the categories are combinations to a total of 105 categories. The number of search patterns is 3223, disambiguation is implemented.

3.2 Lasswell value dictionary (LVD)

The Lasswell value dictionary (LVD) was first introduced with the General Inquirer/Mew Inquirer. A German version was created by Manual Eisner, but is not publicly available. Both the Harvard IV Psychological dictionary and the LVD are included in the General Inquirer/New Inquirer and contain 11788 search patterns. The categories are not mutually exclusive. The category systems are available as Excel file of about 2 MB in size.

3.3 RID – Regressive Imagery Dictionary

The Regressive Imagery Dictionary (Martindale 1975, 1990) is a category system designed to measure primordial vs. conceptual thinking. Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving, whereas primordial thought is associative, concrete, and takes little account of reality. The Regressive Imagery Dictionary contains 65 categories and 3939 search patterns (English version) and 4577 (German): These categories were derived from the theoretical and empirical literature on regressive thought (Martindale 1975, 1990). The rationale behind the dictionary is that psychological processes will be reflected in the content of a text. The RID Regressive Imagery Dictionary is available for several languages and can be used by several programs.

3.4 MCCA – Minnesota Contextual content analysis

MCCA is a general category system with 116 categories for English. The categories of the MCCA are mutually exclusive, and a disambiguation of words takes place. The context is measured with so-called C-scores: traditional, practical, emotional, analytic, the C-scores are used for disambiguation. The number of search patterns is unknown. Nmerical results are computed and compared with standards. The current software version is MCCA 8.3 and runs under Windows.

3.5 LIWC – Linguistic Word Count

The LIWC (Linguistic Inquiry and word count) dictionary consists of 74 variables, 68 of these are non-exclusive categories. The dictionary for English is composed of 2,290 search patterns (words or word stems). Each search pattern defines one or more word categories or subdictionaries. For example, the word 'cried' is part of four word categories: sadness, negative emotion, overall affect, and a past tense verb. Hence, if it is found in the target text, each of these four subdictionary scale scores will be incremented. There are versions for English, Spanish, Italian, German and Dutch. The category system can be used either with LIWC or with WordStat or TextQuest. The Dutch version has 66 categories and 6568 search patterns, the Italian version has 69 Categories and 2829 search patterns, the German version has 49 categories and 5831 search patterns, and the Spanish version has 68 categories and 7460 search patterns.

3.6 Verbal tone of (political) messages

The verbal tone of (political) messages is part of the program Diction. This category system contains 5 main and 35 sub categories, counts them in text blocks and compares the results with built-in standard results. The number of search patterns is approx. 10000, and they cannot be displayed. It is possible to add up to 10 own categories, but the number of search patterns is limited to 200. Characters apart from those in the English alphabet crash the program. The main categories are:

- certainty: Language indicating resoluteness, inflexibility, and completeness and a tendency to speak ex-cathedra.
- activity: Language featuring movement, change, the implementation of ideas and the avoidance of inertia.
- optimism: Language endorsing some person, group, concept or event or highlighting their positive entailments
- commonality: Language highlighting the agreed-upon values of a group and rejecting idiosyncratic modes of engagement.
- realism: Language describing tangible, immediate, recognizable matters that affect people's everyday lives

3.7 Gottschalk-Gleser anxiety scales

PCAD 2000 applies the Gottschalk-Gleser content analysis scales (which measure the magnitude of clearly defined and categorized mental or emotional states) to texts. In addition to deriving scores on a variety of scales, including anxiety, hostility, social alienation, cognitive impairment, hope, and depression, the program compares scores on each scale to norms for the demographic groups of subjects. It can also explain the significance and clinical implications of scores and place subjects into clinical diagnostic classifications derived from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV), developed by the American Psychiatric Association.

The Gottschalk-Gleser anxiety scales are built-in in PCAD 2000. The categories are:

- anxiety (including death, mutilation, separation, guilt, shame, and diffuse anxiety subscales)
- hostility outward (including overt hostility, covert hostility, and total hostility outward subscales)
- hostility inward
- ambivalent hostility (hostility originating externally and directed towards the self)
- social alienation-personal disorganization
- cognitive impairment
- hope
- depression (with 7 subscales)
- health/sickness (with health and sickness subscales)
- human relations
- achievement strivings
- dependency strivings
- quality of life.

3.8 DAL – Dictionary of Language in Affect

The Dictionary of Affect in Language (DAL) consists of 10 scores and is implemented as a macro for Word 2000. Another version was created by Robert Hogenraad to run under Protan. The categories are:

- pleasantness
- activation
- imagery
- rare words
- negations
- use of to be
- colon
- periods
- punctuations
- word length

One of the applications is authorship detection, and DAL compares the results with standard text data.

3.9 Forest value dictionary

This specific dictionary consists of four categories with 615 search patterns and is designed to measure forest values like economic/utilitarian, life support, aesthetic, and moral/spiritual values. It is developed for Wordstat.

3.10 HKW - Hamburger kommunikationssoziologisches Wörterbuch

The HKW (Hamburger kommunikationssoziologisches Wörterbuch – Hamburg’s communication-sociological dictionary) is a general category system for German newspaper headlines. It was published in 1973 for the EVA program (electronic verbal analysis) by Alexander Deichsel. It consists of 86 categories (inspired by the Harvard Psychological Dictionary) and 5872 search patterns. The current version was ported for TextQuest, but some of the categories of the HKW are based on search patterns that are no valid indicators any more (e.g. country names like Jugoslavia, GDR, Soviet Union) or the name of politicians.

3.11 DAW - Dresdner Angstwörterbuch

The DAW (Dresdner Angstwörterbuch – Dresden anxiety dictionary) is based on 6 categories and 3978 search patterns. Originally the category system was developed for CoAn, but it is file compatible with Intext and TextQuest also. An optimised version named DAW2003 reduces the number of search patterns to 1493. The categories are:

- fear of death
- fear of physical damage
- fear of separation
- fear of guilt
- fear of shame, disgrace
- unspecific fear

3.12 Umweltschutzwörterbuch – environmental dictionary

This dictionary was originally created for an agenda-setting study. The category system contains only one category with 107 search patterns, measuring the occurrence of environmental subsections like pollution. It was constructed on the basis of the issue for blind people of two German magazines *Stern* and *Zeit* from 1968-1980. This text contains about 24 MB of text and 150000 different words. The category system is all uppercase and available for Textpack and TextQuest.

3.13 Personal ads

This category system consist of 38 categories for the analysis of personal advertisements in German newspapers. It was used to analyse about 2000 advertisements. The text and the category system are included in TextQuest. The category system consists of 38 different categories with 1363 search patterns, some of them word co-occurrences.

3.14 Television news factors of German television news

This category system consists of 16 news factors and 889 search patterns. The study was conducted in 1986 and found differences in the news factors of public and commercial television. The category system is in German and is part of TextQuest.

4 Software for computer aided content analysis

An overview on available software for dictionary based computer aided content analysis is given on <http://www.textanalysis.info>. There are two free programs: VB-pro by M. Mark Miller, a retired professor of journalism, and INTEXT by Harald Klein. Both run under MS-DOS and include full documentation as a file.

The General Inquirer is a mainframe program that is distributed by ZUMA in Mannheim, Germany. Phil Stone however developed a Java based New Inquirer that processes the Harvard and Lasswell dictionaries stored in an Excel spreadsheet.

The DAL is based on Word macros that are freely available, there is also a Protan version.

Protan is developed by Robert Hogenraad who recently retired, the program is not available for the public and difficult to run. It runs under MS-DOS and Unix. It contains a lot of dictionaries including the RID in all languages, Harvard IV, and a few others.

MCCAlite, Diction, and PCAD2000 come with built-in dictionaries that limit these programs to a specific text genre and language.

LIWC is developed by James W. Pennebaker and runs under Windows and MacOS. The dictionary is available for English, Spanish, German, Italian, Dutch, and Korean.

CoAn is developed by Matthias Romppel and is completely in German. It is an early Windows implementation of INTEXT, the input files are compatible.

Textpack was inspired by the General Inquirer and is developed and distributed by ZUMA in Mannheim, Germany. It runs under MS-Windows and on other platforms where a Fortran compiler is available. Textpack is not equipped with standardised category systems but with some sample files.

TextQuest is the windows version of INTEXT and comes with standardised category systems, mostly for English and German. Text and dictionary files can be exchanged with CoAn and INTEXT.

WordStat is an add-on to Simstat statistical software and runs under Windows. There are some category systems, general and specific ones, provided, and also big thesauri like WordNet and Roget's thesaurus.

There is no general format for a category system, so porting category systems to other programs must be either done manually or with a converting program. Harald Klein wrote three of them for converting Textpack, Protan, and LIWC dictionaries to TextQuest format.

The following table shows which category system works with which software.

Note: WordStat also provides WordNet (over 70000 search patterns) and Roget's Thesaurus (over 100000). CoAn, Intext, and TextQuest are file compatible.

Figure 2: Availability of category systems for different software

category system	GI/NI	LIWC	Protan	TextQuest	WordStat	other
Harvard IV	yes	no	yes	yes	no	no
Lasswell value	yes	no	yes	no	no	no
RID	no	no	yes	yes	yes	no
MCCA	no	no	no	no	no	MCCAlite
LIWC	no	yes	no	yes	yes	no
verbal tone	no	no	no	no	no	Diction
anxiety scales	no	no	no	no	no	PCAD
DAL	no	no	yes	no	no	Word macros
Forest value	no	no	no	no	yes	no
DAW	no	no	no	yes	no	CoAn, Intext
HKW	no	no	no	yes	no	CoAn, Intext
ecological	no	no	no	yes	no	CoAn, Intext, Textpack
personal ads	no	no	no	yes	no	CoAn, Intext
news factors	no	no	no	yes	no	CoAn, Intext
total	2	1	4	8	3	

5 Conclusion

The need for standardised category systems is obvious, and is met by

- programs that contain a sophisticated category system, often with disambiguation, but require a special software to run the category system. The change of the category systems is often not possible or difficult. The programs are mostly language specific and text genre specific.
- programs that are delivered with category systems that can be easily edited to the researcher's need. These programs are sometimes language or text genre specific.

The researcher has to decide which of the category systems is appropriate for the project, and then to choose the software that can run the category system.

If more than one category system is needed, the choice is mostly between TextQuest and WordStat: both programs come with several standardised and specific category systems, and both are not limited to a specific language or text genre. Also changing category systems or building new ones is easy. Wordstat has advanced dictionary construction tools, especially for English, whereas TextQuest has several coding control features including interactive coding of potentially ambiguous and/or negated search pattern and log files.

6 Bibliography

Berth, Hendrik.; Romppel, Matthias. (2000): The measurement of anxiety and dogmatism with computer aided content analysis. In: Blasius, Jörg.; J. Hox; Edith. de Leeuw; Peter Schmidt (Eds.): *Social Science Methodology in the New Millennium. Proceedings of the Fifth International Conference on Logic and Methodology*, Cologne, October 3-6, 2000. TT-Publikaties (CD-ROM, ISBN 90-801073-8-7).

Deichsel, Aleander (1975): *Elektronische Inhaltsanalyse*. Berlin.

Dohrendorf, Rüdiger (1990): Zum publizistischen Profil der *Frankfurter Allgemeinen Zeitung*. Computerunterstützte Inhaltsanalyse von Kommentaren der FAZ. Frankfurt/M, Bern, New York, Paris.

Giegler, Helmut (1991): Zur computerunterstützten Analyse sozialwissenschaftlicher Textdaten. Quantitative und qualitative Strategien. In: Hoffmeier-Zlotnik, Jürgen (eds.): *Analyse qualitativer sozialwissenschaftlicher Daten*. Opladen: 335-388.

Giegler, Helmut; Harald Klein (1994): Correspondence Analysis of Text Data with INTEXT/PC. In: Greenacre, Michael; Jörg Blasius (eds.): *Correspondence Analysis in the Social Sciences*. London: 283-301.

Gottschalk, Louis A. (1995). *Content analysis of verbal behavior: New findings and clinical applications*. Hillsdale, NJ: Lawrence Erlbaum.

Hogenraad, R.; Oriane, E. (1986). Imagery, regressive thinking, and verbal performance in internal monologue. *Imagination, Cognition, and Personality*, 5(2), 127-145.

Klein, Harald (1992): Validity Problems and their Solutions in Computer-Aided Content Analysis with INTEXT/PC and Other New Features. In: Faulbaum, Frank; Reinhold Haux; Karl-Heinz Jöckel (eds.): *Advances in Statistical Software 3*. Stuttgart: 483-388.

Klein, Harald (1996): *Computerunterstützte Inhaltsanalyse mit INTEXT*. Münster.

Martindale, Colin (1975). *Romantic progression: The psychology of literary history*. Washington, D.C.

Martindale, Colin (1990). *The clockwork muse: The predictability of artistic change*. New York.

Namenwirth, J. Zvi (1973): Wheels of time and the Interdependence of Value Change in America. In: *Journal of Interdisciplinary History* 3, 649-683.

Neuendorf, Kimberly A. (2002). *The content analysis handbook*. Thousand Oaks, CA:

Pennebaker, James W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8, 162-166.

Popping, Roel (1999). *Computer-assisted text analysis*. Thousand Oaks, CA: Sage.

Stone, Philip J., Dexter C. Dunphy; Marshall S. Smith, and Daniel M. Ogilvie (1966). *General Inquirer: A computer approach to content analysis*. Cambridge, MA.

Stone, Philip J. and Cambridge Computer Associates Inc. (1968): *User's Manual for the General Inquirer*. Cambridge, MA.

Weber, Robert Philip (1978): *The Dynamics of Value Change, Transformations and Cycles: British speeches from the Throne, 1689-1972*. University of Connecticut.