A Tabular Survey of Automated Table Processing*

Daniel Lopresti[1] and George Nagy[2]

[1] Bell Labs, Lucent Technologies Inc.
600 Mountain Avenue, Room 2D-447
Murray Hill, NJ 07974
dlopresti@lucent.com
[2] Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
nagy@ecse.rpi.edu

Abstract. Tables are the only acceptable means of communicating certain types of structured data. A precise definition of "tabularity" remains elusive because some bureaucratic forms, multicolumn text layouts, and schematic drawings share many characteristics of tables. There are significant differences between typeset tables, electronic files designed for display of tables, and tables in symbolic form intended for information retrieval. Although most research to date has addressed the extraction of low-level geometric information from scanned raster images of paper tables, the recent trend toward the analysis of tables in electronic form may pave the way to a higher level of table understanding.
Recent research on table composition and table analysis has improved our understanding of the distinction between the logical and physical structures of tables, and has led to improved formalisms for modeling tables. The present study indicates that progress on half-a-dozen specific research issues would open the door to using existing paper and electronic tables for database update, tabular browsing, structured information retrieval through graphical and audio interfaces, multimedia table editing, and platform-independent display.
Although tables are not a conventional format for conveying the primary content of technical papers, here we attempt to subdue our natural garrulity by adopting this genre to communicate what we have to say about tables entirely in tabular form.

* Appears in *Graphics Recognition: Recent Advances*, A. K. Chhabra and D. Dori, editors, volume 1941 of Lecture Notes in Computer Science, Springer-Verlag: Berlin, Germany, 2000.

Table 1. *Motivation and definitions.* The study of tables and forms is gaining momentum because of their suitability for electronic information exchange [36]. In this paper we experiment with tables as a means of conveying information that is usually presented in narrative form. Our tables also serve to illustrate formatting and transformations that can be applied to tables. But the medium is NOT our only message.

Why tables?	Prevalent means of communicating structured data Content may include words, numbers, formulas, even graphics Metadata represented by alignment and rulings Adapted to computerized composition Underlying paradigm for spreadsheets and relational databases Bridge between textual and graphic representations
What is a table?	2-D cell assembly for presenting information Regular, repetitive structure along at least one axis [41] Datatype determined by either horizontal or vertical index
What is a form?	Isothetic layout for collecting information One-to-one mapping between indices and data No implication of regularity [41]
What is table analysis?	Information extraction follows table detection and localization Geometric analysis to isolate cell contents Table structure determined simultaneously If needed, OCR translates cells and headers into symbolic form Interpretation requires understanding context
Rationale for this study	Importance of converting tables from one medium to another Rapid growth of tables in various digital formats Desirability of medium-independent query algorithms Interdependence of table composition and interpretation Advent of new applications that require table interpretation Need for research to address neglected table topics

Table 2. *Table of contents.* Table processing draws on established techniques of both text and graphics image analysis, but also requires new research. Starting with a review of current document image analysis, this study leads to a perspective on the relationship between prospective applications and open research areas.

Table 3. *A document taxonomy.* The objective of image analysis and the kind of ancillary data that can facilitate it depends on the document type. Most current DIA applications require processing only documents of a single type.

Type	*Example*	*DIA Task*	*Ancillary Data*
plain text	Moby Dick, Gettysburg Address	extract correct word order	English lexicon & grammar
newspaper, magazine	NY Times, Vogue	separate and reassemble articles; pointers to illustrations, *tables*	publication-specific format
scholarly & technical text	IEEE-PAMI, Dr. Dobbs Journal	index: author, title, page; pointers to refs, figs, *tables*, footnotes, equations	abbreviations, acronyms, units
formal text	program listing, chess, bridge, cookbook	extract executable, or compilable, form	program, chess syntax
letter, memo, envelope	information request, complaint, reservation	extract routing info; index: sender, date, subject	directories
directory	telephone directory, street index	extract name-attribute pairs	previous edition
structured list	organization chart, table of contents, catalog	recover hierarchy; cross-references	previous edition
business form	order, invoice, subscription, survey, IRS-1040	link field content to DBMS; convert to SGML or XML format	formatted data, DBMS, lexicons, workflow
engineering drawing	assembly or part drawing; isometric	convert to CAD format	part lists, drawing stds
schematic diagram	circuits, utility maps	extract net list or convert to CAD format	P-SPICE, cable inventory
map	topographic quad, street map, road map	convert to GIS format	gazetteer, other maps, GIS
table	*airline schedules, stock quotes*	*construct formal model: headers* $\leftrightarrow$ *entries*	*airline, stock abbreviations, previous edition*

Table 4. *Example of a table operation.* The manipulation of rows and columns is a common requirement. The transformation of Table 3 that is illustrated here alters the table to focus attention of the presence of tables in most types of documents. Some documents ("ISA") are best viewed in their entirety as tables or forms.

Type	*Example*	*Tabular Content*
plain text	Moby Dick, Gettysburg Address	none
newspaper, magazine	NY Times, Vogue	stock quotes, temperatures
scholarly & technical text	IEEE-PAMI, Dr. Dobbs Journal	quantitative information
formal text	program listing, chess, bridge, cookbook	repetitive items
letter, memo, envelope	information request, complaint, reservation	delivery schedule, price lists
directory	telephone book, street index	name-attribute pairs
structured list	organization chart, table of contents, catalog	ISA
business form	order, invoice, subscription, survey, IRS-1040	ISA
engineering drawing	assembly or part drawing, isometric	title block, revisions
schematic diagram	circuits, utility maps	component values
map	topo quad, street map, road map	legend
table	*airline schedules, stock quotes*	*ISA*

Table 5. *Common operations in document image analysis.* Tables are in a sense intermediate between mostly-text and mostly-graphics documents. It is therefore instructive to consider the methods of image analysis that have been found useful in these better-established applications. They are organized here bottom-to-top, with the output of the lower-level operations serving for input to the higher-level operations.

	Document Type	
Process Level	*Mostly-text*	*Mostly-graphics*
Pixels	Preprocessing Representation Noise reduction Binarization Skew detection, zoning Character segmentation Script, language, font rec'n Character scaling	Preprocessing Representation Noise reduction Binarization Thinning Vectorization
Primitives	Glyph recognition CC's, strokes Characters, diacritics, punctuation Words	Primitive recognition Straight-lines, curve segments Junctions and nodes Loops Characters
Structures	Text recognition Word segmentation Text line reconstruction Table analysis Morphological content Lexical context Syntax, semantics	Structure recognition Text fields Legends Label attribution Dimensions Graphics symbols Aerial and texture features Beautification (constraints)
Documents	Page layout analysis Text/non-text Physical components Logical components Functional components Compression	Interpretation Component recognition Connectivity analysis CAD/GIS layer separation Database attribute extraction Compression
Corpus	Information retrieval Indexing Search Security, authentication, privacy	DBMS, CAD, GIS interface Validation Update Search

Table 6. *A second example of a table operation.* Condensing the contents of cells and collapsing cell boundaries is useful for accessing tabular information with small displays (palm tops, cell phones). A very small display is illustrated in Fig. 3. A condensed version of Table 5 is shown below.

	Document Type	
Process Level	*Mostly-text*	*Mostly-graphics*
Pixels	Preprocessing	Preprocessing
Primitives	Glyph recognition	Primitive recognition
Structures	Text recognition	Structure recognition
Documents	Page layout analysis	Interpretation
Corpus	Information retrieval	DBMS, CAD, GIS interface

Table 7. *Abstraction in table processing.* As in the case of other types of documents (Tables 5 and 6), the interpretation of tables can be considered at several levels of abstraction. The lowest (image) level is absent in tables prepared for digital media.

Level	*Elements*
Image	pixels
Morphology	geometry: grid, rules, spacing: characters
Syntax	2-D hierarchy; Wang model [51]; text
Semantics	relational data base; natural language processing
Pragmatics	update, retrieval

Table 8. *Growth of table papers.* A simple table that needs lots of context for interpretation! The recent increase in the accessibility of tables in electronic form may be responsible for the sharp growth of table-oriented research.

Years	*# of Pubs*
$\leq$ 1989	11
1990-94	14
1995-98	35

Table 9. *Table papers in the literature.* Relatively few papers attempt to extract semantic information ("content tags").

Analysis	Scanned image	Geometry	[2], [3], [4], [5], [1], [6], [8], [9], [18], [19], [20], [17], [26], [28], [32], [39], [40]
		Cell content analysis	[7], [21], [23], [30], [34], [43], [45], [46], [49], [52], [57]
	Coded text		[25], [29], [42], [41] [13], [24], [27]
Synthesis	Computer		[31], [33], [35], [51]
	Traditional		[10], [22], [48], [54], [55], [56]
Tools	Spreadsheet Database Agents NLP Speech		[37], [38] [14] [16], [45] [13] [12], [44], [47], [50]
Applications	Federal Register Wall Street Journal email		[15] [42] [53]

Table 10. *Table jargon.* Items in Boxhead and Stub are also called Headers, Headings, Labels, Spanning labels, Indices, Captions. There are many books on preferred typesetting practices for tables (see "Traditional" in Table 9). For instance, it was recommended that double-rulings be printed in two passes to avoid gaps at corners.

Stub header	← Boxhead →				
↑			↖	↑	↗
Stub	Cell		←	Block	→
↓			↙	↓	↘

Table 11. *Tables can be recursive.* However, by convention subdivisions increase from top to bottom, and from left to right.

<table>
<tr><td>Tables</td><td>can be</td></tr>
<tr><td>recursive</td><td>
<table>
<tr><td>Tables</td><td>can be</td></tr>
<tr><td>recursive</td><td>
<table>
<tr><td>Tables</td><td>can be</td></tr>
<tr><td>recursive</td><td>
<table>
<tr><td>Tables</td><td>can be</td></tr>
<tr><td>recursive</td><td>...</td></tr>
</table>
</td></tr>
</table>
</td></tr>
</table>
</td></tr>
</table>

Table 12. *Table-form documents.* "Table" and "form" are sometimes used interchangeably, but a clear distinction exists.

Tables	*Forms*
For output	For input
Frame and content created simultaneously	Frame created before content
Tabular structure	Rectilinear structure
Machine-printed	Machine- or hand-printed
Sometimes unique	Frame rarely unique, content often unique

Table 13. *Table representation.* Note: low level can be displayed, intermediate level can be edited, high level can be queried. XML encoding is gaining ground for forms used in commercial transactions, but it is not clear how easy it is to encode meaningfully tables intended for wider use in less specific contexts.

Level of Representation		
Low *("morphology")*	*Intermediate* *("syntax")*	*High* *("semantics")*
PNM/PBM GIF TIF (CCITT, JBIG) PostScript PDF	Rich Text Format Troff, LaTeX HTML MS Word, Excel MatLab Wang Model	Relational DBMS ODA SGML XML

Table 14. *Level of representation.* Rotation is another example of a useful operation. The ordering by level of abstraction is more obvious here than in Table 13.

Level	*Representation*
Low[1] ("morphology")	PNM/PBM, GIF, TIF (CCITT, JBIG), PostScript, PDF
Intermediate[2] ("syntax")	Rich Text Format, Troff, LaTeX, HTML, MS Word, Excel, MatLab, Wang Model
High[3] ("semantics")	Relational DBMS, ODA, SGML, XML

[1] Can be displayed.
[2] Can be edited.
[3] Can be queried.

Table 15. *The Genetic Code I.* Wang [51] developed an abstract data type for tables. It is essentially a forest where each node, except the leaves, are categories called "labeled domains." The categories can be nested. The leaves are the cell contents. The concept of labeled domains is similar to the Dewey Decimal System for library catalogues. In the example below, there are three trees, corresponding to the first, second, and third positions in the genetic code. The entries are amino acids. Each amino acid is specified by the three category labels. In a more complex table, each entry would be specified by a set of "root-to-frontier" paths through the category trees.

Codon Position				*Codon Position*			
1st	*2nd*	*3rd*	*Amino Acid*	*1st*	*2nd*	*3rd*	*Amino Acid*
U	U	U	Phenylalanine	A	U	U	Isoleucine
U	U	C	Phenylalanine	A	U	C	Isoleucine
U	U	A	Leucine	A	U	A	Isoleucine
U	U	G	Leucine	A	U	G	Methionine
U	C	U	Serine	A	C	U	Threonine
U	C	C	Serine	A	C	C	Threonine
U	C	A	Serine	A	C	A	Threonine
U	C	G	Serine	A	C	G	Threonine
U	A	U	Tyrosine	A	A	U	Asparagine
U	A	C	Tyrosine	A	A	C	Asparagine
U	A	A	Stop	A	A	A	Lysine
U	A	G	Stop	A	A	G	Lysine
U	G	U	Cysteine	A	G	U	Serine
U	G	C	Cysteine	A	G	C	Serine
U	G	A	Stop	A	G	A	Arginine
U	G	G	Tryptophan	A	G	G	Arginine
C	U	U	Leucine	G	U	U	Valine
C	U	C	Leucine	G	U	C	Valine
C	U	A	Leucine	G	U	A	Valine
C	U	G	Leucine	G	U	G	Valine
C	C	U	Proline	G	C	U	Alanine
C	C	C	Proline	G	C	C	Alanine
C	C	A	Proline	G	C	A	Alanine
C	C	G	Proline	G	C	G	Alanine
C	A	U	Histidine	G	A	U	Aspartic acid
C	A	C	Histidine	G	A	C	Aspartic acid
C	A	A	Glutamine	G	A	A	Glutamic acid
C	A	G	Glutamine	G	A	G	Glutamic acid
C	G	U	Arginine	G	G	U	Glycine
C	G	C	Arginine	G	G	C	Glycine
C	G	A	Arginine	G	G	A	Glycine
C	G	G	Arginine	G	G	G	Glycine

Table 16. *The Genetic Code II.* Wang calls the number of categories the "dimension" of the table. The Genetic Code is three-dimensional, regardless of its physical layout. In the rendering below, the cells are arranged to minimize the repetition of cell entries. The "size" of a table is the product of the number of lowest-level categories, here $4 \times 4 \times 4 = 64$.

UUU	Phenyl-	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
UUC	alanine	UCC		UAC		UGC	
UUA	Leucine	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Tryptophan
CUU		CCU	Proline	CAU	Histidine	CGU	Arginine
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Glutamine	CGA	
CUG		CCG		CAG		CGG	
AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lysine	AGA	Arginine
AUG	Methionine	ACG		AAG		AGG	
GUU	Valine	GCU	Alanine	GAU	Aspartic	GGU	Glycine
GUC		GCC		GAC	acid	GGC	
GUA		GCA		GAA	Glutamic	GGA	
GUG		GCG		GAG	acid	GGG	

Table 17. *The Genetic Code III.* Here the first and third categories are laid out vertically, and the second category horizontally. Many other possible permutations exist. Wang also developed software for creating different tabular layouts for the same logical table. She found that most of the several hundred tables in standard texts and monographs that she examined fit her model, except for the frequent presence of footnotes. Wang's main contribution is the separation between the logical and physical aspects of a table.

First	*Second Position*				*Third*
Position	U	C	A	G	*Position*
	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Table 18. *Strategies for extracting table geometry.* (Issues: Hierarchical vs. flat structure? Skew invariance? Start with cells or with external frame?)

		Model-driven		*Data-driven*	
		Top-down	*Bottom-up*	*Top-down*	*Bottom-up*
Primitives	Rulings	√	√	√	√
	White space	√	√		
	Text	√	√	√	√
	Cell		√		√

Table 19. *Logical/functional analysis.* In contrast to the data-driven analysis described in Table 18, here the analysis is model-driven.

Table syntax	Green and Krishnamoorthy [19, 18, 20]
Structure description tree	Watanabe, Quo, and Sugie [52]
Cohesion domain template	Hurst [27]
OSM	Embley, Kurtz, and Woodfield [14], Haas [21]
Abstract data type	Wang [51]
Relational algebra	Codd [11]

Table 20. *Some sources of difficulty.* The Appendix has examples that illustrate many of the problems that would have to be solved in developing a broad-gauge table-understanding system. Note, however, that none of the example tables are particularly difficult from the standpoint of human perception, though some require either specialized knowledge (Figs. 5 and 9) or the appropriate mindset (Figs. 12 and 13).

Morphology	Violations of tabular layout
	Incomplete gird rulings
	Close-spaced or misaligned cells
	Misplaced or oddly-oriented headers
	Multi-text-line cells
Syntax	Multi-dimensional structure
	Unusual layout
	Combined tables
	Split tables
	Footnotes[4]
Semantics	OCR or other errors in text
	Synonyms, abbreviations
	Incomplete headers
	Missing data-definition dictionary
	Iconic cell contents

[4] Wang surveyed nearly 900 tables and found that 40% contain footnotes [51], pg. 154.

Table 21. *Applications and research problems.* We have identified several classes of potential applications for table processing and some research problems on which little work has been reported so far. We have also formed opinions of the relative difficulties of the tasks involved. The ways in which the applications and problems interrelate are depicted below. Unless we make headway on performance evaluation, including acquisition of statistically adequate test material, it will be difficult to evaluate progress on any of the other tasks.

	Query mechanisms	Audio navigation	Table subdivision	Table spotting	Table clustering	Conversion to abstract form	Overcoming recognition errors	Performance evaluation
Large-volume, homogeneous conversion						•	•	•
Large-volume, mixed conversion				•	•	•	•	•
Individual database creation	•		•	•	•	•	•	•
Tabular browsing	•	•	•	•	•	•	•	•
Audio access to tables	•	•	•	•	•	•	•	•
Table manipulation			•		•	•	•	•
Table modification for display			•		•	•	•	•

References

1. A. Abu-Tarif. Table processing and table understanding. Master's thesis, Rensselaer Polytechnic Institute, May 1998.
2. J. F. Arias, S. Balasubramanian, A. Prasad, R. Kasturi, and A. Chhabra. Information extraction from telephone company drawings. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 729–732, Seattle, Washington, June 1994.
3. J. F. Arias, A. Chhabra, and V. Misra. Efficient interpretation of tabular documents. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume III, pages 681–685, Vienna, Austria, August 1996.
4. J. F. Arias, A. Chhabra, and V. Misra. Interpreting and representing tabular documents. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 600–605, San Francisco, CA, June 1996.
5. J. F. Arias and R. Kasturi. Efficient techniques for line drawing interpretation and their application to telephone company drawings. Technical Report CSE TR CSE-95-020, Penn State University, August 1995.
6. S. Balasubramanian, S. Chandran, J. F. Arias, R. Kasturi, and A. Chhabra. Information extraction from tabular drawings. In *Proceedings of Document Recognition I*

(IS&T/SPIE Electronic Imaging'94), volume 2181, pages 152–163, San Jose, CA, June 1994.

7. L. Bing, J. Zao, and X. Hong. New method for logical structure extraction of form document image. In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging'99)*, volume 3651, pages 183–193, San Jose, CA, January 1999.
8. S. Chandran and R. Kasturi. Structural recognition of tabulated data. In *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 516–519, Tsukuba Science City, Japan, October 1993.
9. A. K. Chhabra, V. Misra, and J. Arias. Detection of horizontal lines in noisy run length encoded images: The FAST method. In R. Kasturi and K. Tombre, editors, *Graphics Recognition – Methods and Applications*, volume 1072 of *Lecture Notes in Computer Science*, pages 35–48. Springer-Verlag, Berlin, Germany, 1996.
10. *The Chicago Manual of Style*. The University of Chicago Press, 1982.
11. E. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), June 1970.
12. M. J. DeHaemer, G. Wright, and T. W. Dillon. Automated speech recognition for spreadsheet tasks: Performance effects for experts and novices. *International Journal of Human-Computer Interaction*, 6(3):299–318, 1994.
13. S. Douglas, M. Hurst, and D. Quinn. Using natural language processing for identifying and interpreting tables in plain text. In *Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, pages 535–545, Las Vegas, NV, April 1995.
14. D. Embley, B. Kurtz, and S. Woodfield. *Object-oriented Systems Analysis: A Model Driven Apprach*. Yourdon Press, 1992.
15. M. Garris, S. Janet, and W. Klein. Federal Register document image database. In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging'99)*, volume 3651, pages 97–108, San Jose, CA, January 1999.
16. P. Gray, S. Embury, W. Gray, and K. Hui. An agent-based system for handling distributed design constraints. In *Proceedings of Agents'98*, 1998.
17. E. A. Green. *Model-based analysis of printed tables*. PhD thesis, Rensselaer Polytechnic Institute, May 1996.
18. E. A. Green and M. Krishnamoorthy. Model-based analysis of printed tables. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, pages 214–217, Montréal, Canada, August 1995.
19. E. A. Green and M. Krishnamoorthy. Model-based analysis of printed tables. In *Proceedings of the First International Workshop on Graphics Recognition (GREC'95)*, pages 234–242, PA, 1995.
20. E. A. Green and M. Krishnamoorthy. Recognition of tables using table grammars. In *Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, pages 261–277, Las Vegas, NV, April 1995.
21. T. B. Haas. The development of a prototype knowledge-based table-processing system. Master's thesis, Brigham Young University, December 1997.
22. R. Hall. *Handbook of Tabular Presentation*. The Ronald Press Company, New York, NY, 1943.
23. Y. Hirayama. A method for table structure analysis using DP matching. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, pages 583–586, Montréal, Canada, August 1995.
24. O. Hori and D. S. Doermann. Robust table-form structure analysis based on box-driven reasoning. In *Proceedings of the Third International Conference on Doc-*

ument Analysis and Recognition (ICDAR'95), pages 218–221, Montréal, Canada, August 1995.
25. J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Medium-independent table detection. In *Proceedings of Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging'00)*, San Jose, CA, January 2000. To appear.
26. T. Hu. Recognizing table entries in a scanned document. Master's thesis, Rensselaer Polytechnic Institute, October 1993.
27. M. Hurst and S. Douglas. Layout and language: Preliminary investigations in recognizing the structure of tables. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, pages 1043–1047, August 1997.
28. K. Itonori. A table structure recognition based on textblock arrangement and ruled line position. In *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 765–768, Tsukuba Science City, Japan, October 1993.
29. T. G. Kieninger. Table structure recognition based on robust block segmentation. In *Proceedings of Document Recognition V (IS&T/SPIE Electronic Imaging'98)*, volume 3305, pages 22–32, San Jose, CA, January 1998.
30. W. Kornfeld and J. Wattecamps. Automatically locating, extracting and analyzing tabular data. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–348, Melbourne, Australia, August 1998.
31. M. Krishnamoorthy. TBL, an easy to use table description language. Internal document, Rensselaer Polytechnic Institute, 1992.
32. G. Kyriazis. Analysis of digitized tables. Senior project report, Rensselaer Polytechnic Institute, 1990.
33. L. Lamport. *LaTeX: A Document Preparation System.* Addison-Wesley, Reading, MA, 1985.
34. A. Laurentini and P. Viada. Identifying and understanding tabular material in compound documents. In *Proceedings of the Eleventh International Conference on Pattern Recognition (ICPR'92)*, pages 405–409, The Hague, 1992.
35. M. Lesk. Tbl – a program to format tables. In *UNIX Programmer's Manual*, volume 2A. Bell Telephone Laboratories, Murray Hill, NJ, 1979.
36. D. Lopresti and G. Nagy. Automated table processing: An (opinionated) survey. In *Proceedings of the Third IAPR International Workshop on Graphics Recognition*, pages 109–134, Jaipur, India, September 1999.
37. *Lotus 1-2-3 User's Handbook.* Ballantine Books, New York, NY, 1984.
38. *Microsoft Excel User's Guide.* Microsoft Corporation, Redmond, WA, 1990.
39. G. Nagy, M. Krishnamoorthy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):737–747, 1993.
40. G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Proceedings the International Conference on Pattern Recognition (ICPR)*, pages 347–349, 1984.
41. C. Peterman, C. H. Chang, and H. Alam. A system for table understanding. In *Proceedings of the Symposium on Document Image Understanding Technology (SDIUT'97)*, pages 55–62, Annapolis, MD, April/May 1997.
42. P. Pyreddy and W. B. Croft. TINTIN: A system for retrieval in text tables. Technical Report UM-CS-1997-002, University of Massachusetts, Amherst, January 1997.

43. M. A. Rahgozar and R. Cooperman. A graph-based table recognition system. In *Proceedings of Document Recognition III (IS&T/SPIE Electronic Imaging'96)*, volume 2660, pages 192–203, San Jose, CA, January 1996.
44. *The 1.7 Tag Set Usage Guide*. Recording for the Blind and Dyslexic, Princeton, NJ, 1994.
45. D. Rus and D. Subramanian. Customizing information capture and access. *ACM Transactions on Information Systems*, 15(1):67–101, 1997.
46. J. H. Shamalian, H. S. Baird, and T. L. Wood. A retargetable table reader. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, pages 158–163, August 1997.
47. R. Sproat, J. Hu, and H. Chen. EMU: an e-mail preprocessor for text-to-speech. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 239–244, Los Angeles, CA, December 1998.
48. E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.
49. E. Turolla, Y. Belaid, and A. Belaid. Form item extraction based on line searching. In R. Kasturi and K. Tombre, editors, *Graphics Recognition – Methods and Applications*, volume 1072 of *Lecture Notes in Computer Science*, pages 69–79. Springer-Verlag, Berlin, Germany, 1996.
50. M. A. Walker, J. Fromer, G. D. Fabbrizio, C. Mestel, and D. Hindle. What can I say?: Evaluating a spoken language interface to email. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 582–589, Los Angeles, CA, April 1998.
51. X. Wang. *Tabular abstraction, editing, and formatting*. PhD thesis, University of Waterloo, 1996.
52. T. Watanabe, Q. L. Quo, and N. Sugie. Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):432–445, 1995.
53. S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 276–283, Vancouver, British Columbia, Canada, April 1996.
54. P. Wright. Using tabulated information. *Ergonomics*, 11(4):331–343, 1968.
55. P. Wright. Understanding tabular displays. *Visible Language*, 7:351–359, 1973.
56. P. Wright. The comprehension of tabulated information: some similarities between prose and reading tables. *NSPI Journal*, XIX(8):25–29, October 1980.
57. K. Zuyev. Table image segmentation. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, pages 705–708, August 1997.

Appendix: Table Examples

In this appendix, we present a number of examples of paper and electronic tables.

TABLE 1
A BRIEF SURVEY OF GEOMETRIC PAGE-LAYOUT ANALYSIS METHODS

No.	Author	Year	Approach	Features
1	Wahl et al. [11]	1982	Run length smoothing	Time consuming and skew sensitive
2	Nagy et al. [12]	1984	X-Y tree cut	Skew sensitive; Assumes rectangular blocks
3	Wang et al. [13]	1989	Run length smoothing and recursive X-Y cut	Newspaper analysis; Sensitive to skew
4	Fujisawa et al. [14]	1990	Top-down	Japanese patent documents
5	Fisher et al. [15]	1990	Run length smoothing and connected component extraction	Identifies text and nontext zones; Skew sensitive
6	Pavlidis et al. [16]	1991	Column oriented projection	Identifies text and nontext regions; Accommodates moderate skew
7	Baird [17]	1992	Global-to-local strategy	Accommodates different languages; Skew correction;
8	Jain et al. [18]	1992	Gabor filtering	Multichannel texture features from gray-scale images; Time consuming
9	Lebourgeois et al. [19]	1992	8× 3 window filtering	Unconstrained documents; Skew not considered
10	Pavlidis et al. [20]	1992	Horizontal smearing and bottom-up	Accommodates small skew; Fixed parameters
11	Akindele et al. [21]	1993	White space tracing	Polygonal blocks; Only text zones considered
12	Amamoto et al. [22]	1993	Morphological operation on white space	Identifies horizontal and vertical writing; Skew not considered
13	Ittner et al. [23]	1993	White space and minimum spanning tree	Language and orientation free; Large computation
14	O'Gorman [24]	1993	*k*-nearest neighbor clustering	Can handle arbitrary orientation with high accuracy; Large computation
15	Antonacopoulos et al. [25], [26]	1994	Contours from white tiles	Finds nonrectangular and skewed regions; Error in classifying large fonts
16	Zlatopolsky [27]	1994	Connected component extraction	Multiple skewed document; Sensitive parameters
17	Doermann [28]	1995	Wavelet multiscale analysis	Segments nonblock-nested pages; Gray-scale image processing; High computational complexity
18	Drivas et al. [29]	1995	Connected component grouping	Skew correction with a time consuming algorithm
19	Ha et al. [30]	1995	Connected component-based projection profile	Faster than pixel-based projection profile; Skew sensitive
20	Sylwester et al. [31]	1995	trainable X-Y cut	Relatively robust; Skew and noise free
21	Tang et al. [32]	1995	Modified fractal signature	Handles documents with high geometrical complexity; Gray-scale image processing; Time consuming
22	Jain et al. [33], [34]	1996	Masks and neural network	Handles documents with multiple languages; Gray-scale image processing; Time consuming
23	Kise et al. [35]	1996	Background thinning	Skewed nonrectangular layout; Bounding box is not very tight
24	Liu et al. [36]	1996	Adaptive top-down and bottom-up	Nonrectangular regions; Skew free
25	Yamashita et al. [37]	1996	Run length smearing and adaptive thresholding	Less sensitive to font size and spacing; Skew free

Fig. 1. A table with considerable text comparing document layout analysis methods.[5] Except for multi-line cells, this table has no irregular features that would complicate analysis. There are three categories: Citation, Method, and "No.", but the first two are implicit at the root level and only evident from the subcategory labels.

[5] From "Document Representation and its Application to Page Decomposition" by A. K. Jain and B. Yu, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 1998, pg. 297.

NAME		ADDRESS					TELNO		
First	Last	#	Street	City	State	Zip	Area-Code	#	Extension
...		...					...		

Fig. 2. Multiple column headers, where the top header subsumes several headers at the next level, are common. This makes it difficult to separate "domains" and "subdomains" (Wang's terminology) for subsequent analysis. Style manuals recommend avoiding horizontal rulings (*The Government Printing Office Style Manual* has over thirty pages of guidelines on "tabular work").

Fig. 3. A very small table.[6] In the scanned image shown, low and irregular contrast would complicate pixel-level analysis. However, the watch is only an example of a small digital display, from which the information would be obtained in computer-readable form rather than by optical scanning. At the logical level, lack of space precludes headers: the only clues are the usual functions of a watch, and the formatting of the entries.

[6] From one of the author's Casio DataBank 150 watch.

																8
																38
																31
															34	34
															86	86
															84	84
															87	87
														40	40	40
														39	39	39
														88	88	88
														81	81	81
														7	7	7
														85	85	85
													36	36	36	36
													91	91	91	91
													79	79	79	79
													76	76	76	76
												29	29	29	29	29
												37	37	37	37	37
												90	90	90	90	90
												92	92	92	92	92
												89	89	89	89	89
												83	83	83	83	83
												78	78	78	78	78
											80	80	80	80	80	80
											30	30	30	30	30	30
											33	33	33	33	33	33
											6	6	6	6	6	6
											27	27	27	27	27	27
											4	4	4	4	4	4
											28	28	28	28	28	28
											35	35	35	35	35	35
											82	82	82	82	82	82
										77	77	77	77	77	77	77
									3	3	3	3	3	3	3	3
									5	5	5	5	5	5	5	5
								32	32	32	32	32	32	32	32	32
						41	41	41	41	41	41	41	41	41	41	41
					42	42	42	42	42	42	42	42	42	42	42	42
			26	26	26	26	26	26	26	26	26	26	26	26	26	26
		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
			1917									1918				

Fig. 4. A table showing the stationing of U.S. Army Divisions in France during WWI.[7] The use of blanks makes this table look like a graph, which complicates extraction of the tabular structure. Note the vestigial ruling between "Dec 1917" and "Jan 1918."

[7] From *The Visual Display of Quantitative Information* by Edward R. Tufte, Graphics Press: Cheshire, CT, 1983, pg. 141.

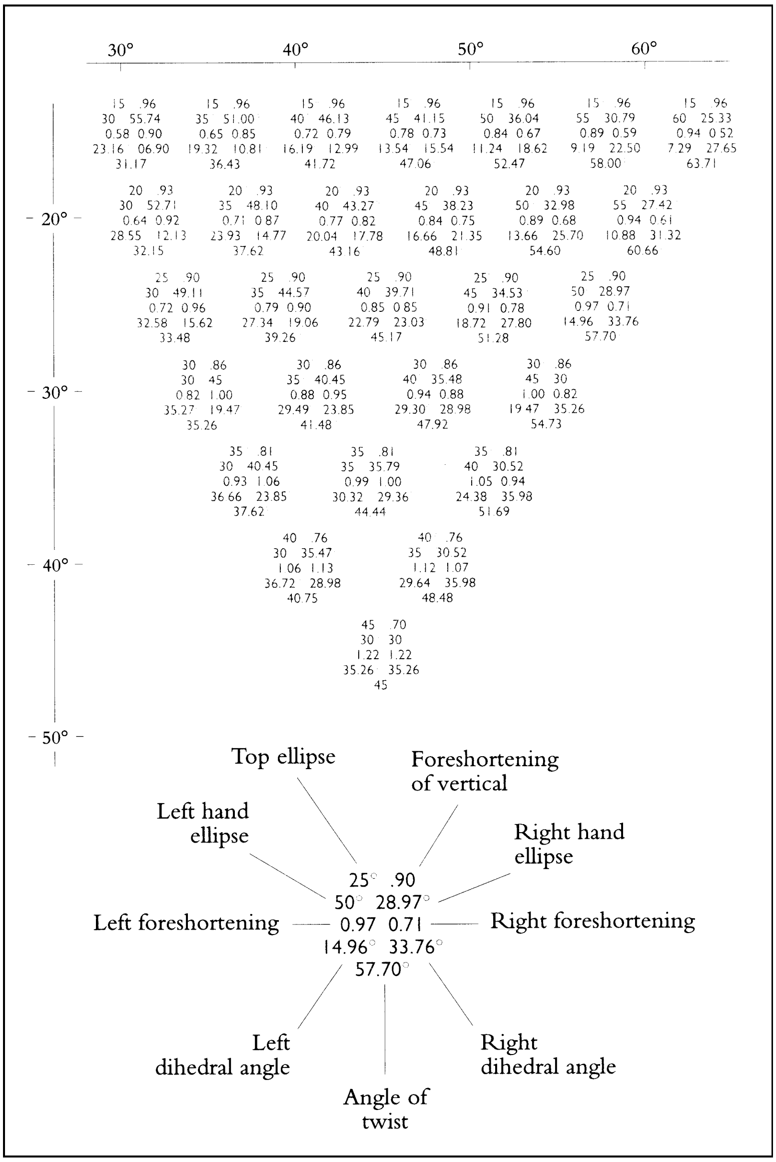


Fig. 5. A table presenting nine parameters for a cube in triametric projection.[8] This table may also be classified as a diagram. The last cell in the third row is recursively expanded in the bottom half. It would be difficult to define the Wang dimensionality of this example because it lacks rectilinear structure.

[8] From *Visual Explanations* by Edward R. Tufte, Graphics Press: Cheshire, CT, 1997, pg. 85.

How Different Groups Voted for President

Based on 12,782 interviews with voters at their polling places. Shown is how each group divided its vote for President and, in parentheses, the percentage of the electorate belonging to each group.

	CARTER	REAGAN	ANDERSON	CARTER-FORD in 1976
Democrats (43%)	66	26	6	77 - 22
Independents (23%)	30	54	12	43 - 54
Republicans (28%)	11	84	4	9 - 90
Liberals (17%)	57	27	11	70 - 26
Moderates (46%)	42	48	8	51 - 48
Conservatives (28%)	23	71	4	29 - 70
Liberal Democrats (9%)	70	14	13	86 - 12
Moderate Democrats (22%)	66	28	6	77 - 22
Conservative Democrats (8%)	53	41	4	64 - 35
Politically active Democrats (3%)	72	19	8	—
Democrats favoring Kennedy in primaries (13%)	66	24	8	—
Liberal Independents (4%)	50	29	15	64 - 29
Moderate Independents (12%)	31	53	13	45 - 53
Conservative Independents (7%)	22	69	6	26 - 72
Liberal Republicans (2%)	25	66	9	17 - 82
Moderate Republicans (11%)	13	81	5	11 - 88
Conservative Republicans (12%)	6	91	2	6 - 93
Politically active Republicans (2%)	5	89	6	—
East (32%)	43	47	8	51 - 47
South (27%)	44	51	3	54 - 45
Midwest (20%)	41	51	6	48 - 50
West (11%)	35	52	10	46 - 51
Blacks (10%)	82	14	3	82 - 16
Hispanics (2%)	54	36	7	75 - 24
Whites(88%)	36	55	8	47 - 52
Female (49%)	45	46	7	50 - 48
Male (51%)	37	54	7	50 - 48
Female, favors equal rights amendment (22%)	54	32	11	—
Female, opposes equal rights amendment (15%)	29	66	4	—
Catholic (25%)	40	51	7	54 - 44
Jewish (5%)	45	39	14	64 - 34
Protestant (46%)	37	56	6	44 - 55
Born-again white Protestant (17%)	34	61	4	—
18 - 21 years old (6%)	44	43	11	48 - 50
22 - 29 years old (17%)	43	43	11	51 - 46
30 - 44 years old (31%)	37	54	7	49 - 49
45 - 59 years old (23%)	39	55	6	47 - 52
60 years or older (18%)	40	54	4	47 - 52
Family income				
Less than $10,000 (13%)	50	41	6	58 - 40
$10,000 - $14,999 (14%)	47	42	8	55 - 43
$15,000 - $24,999 (30%)	38	53	7	48 - 50
$25,000 - $50,000 (24%)	32	58	8	36 - 62
Over $50,000 (5%)	25	65	8	—
Professional or manager (40%)	33	56	9	41 - 57
Clerical, sales or other white-collar (11%)	42	48	8	46 - 53
Blue-collar worker (17%)	46	47	5	57 - 41
Agriculture (3%)	29	66	3	—
Looking for work (3%)	55	35	7	65 - 34
Education				
High school or less (39%)	46	48	4	57 - 43
Some college (28%)	35	55	8	51 - 49
College graduate (27%)	35	51	11	45 - 55
Labor union household (26%)	47	44	7	59 - 39
No member of household in union (62%)	35	55	8	43 - 55
Family finances				
Better off than a year ago (16%)	53	37	8	30 - 70
Same (40%)	46	46	7	51 - 49
Worse off than a year ago (34%)	25	64	8	77 - 23
Family finances and political party				
Democrats, better off than a year ago (7%)	77	16	6	69 - 31
Democrats, worse off than a year ago (13%)	47	39	10	94 - 6
Independents, better off (3%)	45	36	12	—
Independents, worse off (9%)	21	65	11	—
Republicans, better off (4%)	18	77	5	3 - 97
Republicans, worse off (11%)	6	89	4	24 - 76
More important problem				
Unemployment (39%)	51	40	7	75 - 25
Inflation (44%)	30	60	9	35 - 65
Feel that U.S. should be more forceful in dealing with Soviet Union even if it would increase the risk of war (54%)	28	64	6	—
Disagree (31%)	56	32	10	—
Favor equal rights amendment (46%)	49	38	11	—
Oppose equal rights amendment (35%)	26	68	4	—
When decided about choice				
Knew all along (41%)	47	50	2	44 - 55
During the primaries (13%)	30	60	8	57 - 42
During conventions (8%)	36	55	7	51 - 48
Since Labor Day (8%)	30	54	13	49 - 49
In week before election (23%)	38	46	13	49 - 47

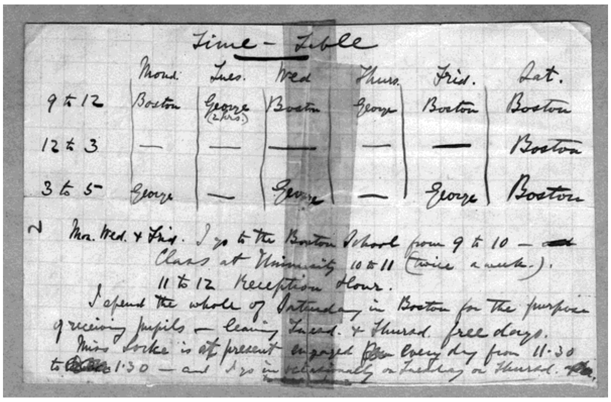
Source: 1976 and 1980 election day surveys by The New York Times/CBS News Poll and 1976 election day survey by NBC News.

Fig. 6. A table analyzing voter preferences in the 1980 U.S. Presidential Election.[9] Some of the category labels, like political affiliation and gender, are implicit. Therefore any automated interpeter would require a built-in understanding of demographic categories.

[9] From *The Visual Display of Quantitative Information* by Edward R. Tufte, Graphics Press: Cheshire, CT, 1983, pg. 179. Tufte notes: "This type of elaborate table, a *supertable*, is likely to attract and intrigue readers through its organized, sequential detail and reference-like quality. One supertable is far better than a hundred little bar charts."

Monday, September 20, 1999			
	Track A Convention Hall A	Track B Convention Hall B	Track C Chanakya Hall
08:30 10:00	OPENING SESSION(Mo-1) Banquet Hall		
10:00	COFFEE BREAK Pool Side		
10:30 12:30	MULTIMEDIA DOCUMENT PROCESSING Mo-2A	CHARACTER RECOGNITION Mo-2B	DOCUMENT IMAGE PROCESSING - I Mo-2C
12:30	LUNCH Pool Side		
13:30 14:30	POSTER PRESENTATION Mo-3A	POSTER PRESENTATION Mo-3B	POSTER PRESENTATION Mo-3C
13:30 15:30	POSTER SESSION-I (Mo-3) Banquet Hall (Coffee served at 14:30)		
15:30 17:30	INFORMATION RETRIEVAL Mo-4A	POSTAL AUTOMATION Mo-4B	FONT RECOGNITION Mo-4C
19:00 21:00	CONFERENCE RECEPTION Banquet Hall		

Fig. 7. ICDAR'99 schedule.[10] This schedule, which was perfectly clear to the conference attendees, has many irregularities to confuse automated analysis. The information in each column may be a title or a location. Times are shown inconsistently on the left. By introducing a cross-track category for social functions, it would be possible to rationalize the structure.



Time - Table

	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
9 to 12	Boston	George (2 hrs)	Boston	George	Boston	Boston
12 to 3	—	—	—	—	—	Boston
3 to 5	George	—	George	—	George	Boston

Mon. Wed. & Fri. I go to the Boston School from 9 to 10 —
Class at University 10 to 11 (twice a week).
11 to 12 Reception Hour.
I spend the whole of Saturday in Boston for the purpose of receiving pupils — leaving Tuesd. & Thursd. free days.
Miss Locke is at present engaged every day from 11.30 to 1.30 — and I go in occasionally on Tuesday or Thursd.

Fig. 8. A handwritten table showing a personal schedule.[12] In handwritten tables like this, both structure extraction and text interpretation are difficult and error-prone. We have seen no work on handwritten tables, but much effort has been devoted to block-lettered tables in engineering drawings and to hand-filled forms. In successful applications a considerable amount of context is available to guide interpretation.

[10] From http://www.cedar.buffalo.edu/ICDAR99/Program/page12.html.

[12] From the Library of Congress archive of the Alexander Graham Bell family papers, http://memory.loc.gov/ammem/bellhtml/bellhome.html.

1996 DDVPC 2400bps Candidate Coder Intellibility Performance Calibrations

	Wgt	Wgt	Wgt	A	A	A	A	B	B	B	B	C	C	C	C	D	D	D	D
DRT	0.35			A(M)	A(F)	A(C)	A(SE)	B(M)	B(F)	B(C)	B(SE)	C(M)	C(F)	C(C)	C(SE)	D(M)	D(F)	D(C)	D(SE)
Quiet		0.20	0.100	90.8	88.9	89.9	0.55	90.1	90.8	90.5	0.48	93.8	90.8	92.3	0.70	92.7	90.1	91.4	0.71
Vinson Quiet			0.100	91.5	91.4	91.4	0.42	91.4	88.4	89.9	0.83	91.3	91.1	91.2	0.78	92.1	89.4	90.8	0.57
Office		0.60	0.067	91.8	91.2	91.5	0.91	88.6	88.2	88.4	0.55	91.5	91.0	91.2	0.75	92.4	89.8	91.1	0.59
Auto			0.067	87.9	77.6	82.7	0.93	86.3	79.2	82.7	0.91	86.8	79.9	83.4	0.81	87.5	81.3	84.4	0.88
Humvee			0.067	61.8	66.9	64.4	1.06	64.8	64.9	64.8	1.30	61.8	64.4	63.1	0.72	61.2	69.6	65.4	0.62
M2 Bradley			0.067	67.4	67.1	67.3	1.36	66.3	68.8	67.6	0.88	61.4	67.2	64.3	1.17	66.0	67.1	66.6	0.71
Helicopter			0.067	61.5	68.5	65.0	1.15	68.6	68.9	68.7	1.07	63.5	69.9	66.7	0.61	66.2	71.2	68.7	1.04
F-15			0.067	74.9	78.1	76.5	0.85	71.1	71.1	71.1	0.52	77.5	76.7	76.7	0.91	73.4	75.2	74.3	0.57
E3A			0.067	87.6	83.0	85.3	0.75	86.9	84.8	85.9	0.81	89.5	85.5	87.5	0.85	89.1	83.8	86.5	0.66
P3C			0.067	89.2	82.2	85.7	0.71	87.4	81.5	84.4	0.75	84.5	85.4	85.4	0.59	87.3	81.9	84.6	0.88
MCE			0.067	88.5	88.4	88.5	0.72	86.7	87.5	87.1	0.77	89.0	90.6	89.8	1.08	86.6	88.9	87.8	0.84
BER		0.10	0.050	86.2	82.6	84.4	0.89	88.5	85.7	87.1	0.52	88.6	89.2	88.9	0.63	87.2	86.1	86.7	0.71
BLER			0.050	92.3	89.7	91.0	0.86	92.4	89.7	91.1	0.63	93.4	90.0	91.7	0.64	90.2	88.2	89.2	0.45
S_Tandem		0.10	0.050	85.0	85.3	85.2	0.73	87.4	84.3	85.9	0.60	87.5	84.5	86.0	0.72	87.6	84.7	86.2	0.71
D_Tandem			0.050	84.1	81.9	83.0	0.86	83.0	78.0	80.5	0.62	85.2	80.6	82.9	0.58	84.3	81.2	82.8	0.75
Intell. Perf				82.984	81.872	82.437	0.261	82.829	81.132	81.984	0.293	83.279	82.779	83.032	0.241	83.259	82.214	82.759	0.217
		1.00	1.000	A(M)	A(F)	A(C)	A(SE)	B(M)	B(F)	B(C)	B(SE)	C(M)	C(F)	C(C)	C(SE)	D(M)	D(F)	D(C)	D(SE)
Rank				4	4	4		5	6	5		2	2	2		3	3	3	

1996 DDVPC 2400bps Reference Coder Intellibility Performance Calibrations

	Wgt	Wgt	Wgt	CELP	CELP	CELP	CELP	CVSD	CVSD	CVSD	CVSD	LPC	LPC	LPC	LPC
DRT	0.35			Celp(M)	Celp(F)	Celp(C)	Celp(SE	KY(M)	KY(F)	KY(C)	KY(SE)	LPC(M)	LPC(F)	LPC(C)	LPC(SE)
Quiet		0.20	0.100	90.9	90.5	90.7	0.34	88.2	88.8	88.5	0.85	87.3	85.1	86.2	0.60
Vinson Quiet			0.100	91.9	90.3	91.1	0.44	91.4	90.2	90.8	0.66	81.9	82.7	82.3	0.65
Office		0.60	0.067	89.8	88.3	89.0	0.88	89.6	88.1	88.8	0.50	84.8	85.5	85.2	0.81
Auto			0.067	88.9	83.3	86.1	0.85	89.0	84.8	86.9	1.02	73.1	63.7	68.4	0.73
Humvee			0.067	60.6	65.4	63.0	0.95	65.2	73.3	69.3	1.33	21.7	41.7	31.7	2.26
M2 Bradley			0.067	60.7	66.9	63.8	1.14	74.3	78.4	76.4	0.94	34.2	42.5	38.4	1.27
Helicopter			0.067	61.0	66.6	63.8	0.94	75.6	78.9	77.2	0.78	39.4	55.8	47.6	1.24
F-15			0.067	73.0	75.5	74.3	0.79	74.7	78.6	76.6	1.11	70.5	69.4	69.9	0.88
E3A			0.067	84.6	85.6	85.0	0.62	88.2	89.9	89.0	0.94	66.7	65.3	66.0	1.09
P3C			0.067	85.7	82.7	84.2	1.19	89.5	86.0	87.7	0.72	80.9	78.5	79.7	1.00
MCE			0.067	90.5	87.8	89.1	0.96	90.8	90.0	90.4	0.75	77.5	78.7	78.1	1.10
BER		0.10	0.050	90.3	86.0	88.2	0.73	86.1	87.8	86.9	0.67	80.0	82.7	81.4	0.90
BLER			0.050	87.1	89.3	88.2	0.96	86.8	85.6	86.2	0.78	85.2	82.9	84.0	0.72
S_Tandem		0.10	0.050	84.8	83.7	84.3	0.61	89.0	87.3	88.2	1.03	75.8	75.5	75.6	1.10
D_Tandem			0.050	83.0	80.6	81.8	0.96	84.4	85.9	85.2	0.70	72.0	73.5	72.7	0.64
Intell. Perf				81.860	81.867	81.859	0.260	84.402	85.097	84.742	0.297	69.157	71.250	70.202	0.332
		1.00	1.000	Celp(M)	Celp(F)	Celp(C)	Celp(SE	KY(M)	KY(F)	KY(C)	KY(SE)	LPC(M)	LPC(F)	LPC(C)	LPC(SE)
Rank				6	5	6		1	1	1		7	7	7	

Fig. 9. A wide, wrapped table giving the performance of various voice coding schemes.[13] The identical leftmost columns and different column headers confirm that this is a split table. Distinctions like that between "Quiet" and "Vinson Quiet" require expert interpretation. The abbreviations are the least of the difficulty, since they could be expanded with table look-up. One of the columns with Rank 2 is selected for special consideration.

[13] From "A New Federal Standard Algorithm for 2400bps Coded Voice." Note the extra, inexplicable (in this context) box surrounding the performance and rank figures for the entry in the middle of the first part of the table. http://www.plh.af.mil/ddvpc/24results.htm.

```
*******************************************************
               LUCENT TECHNOLOGIES TODAY
         For the People of Lucent Technologies
               Friday, February 12, 1999
*******************************************************

                 ***  STOCK WATCH  ***

               TODAY'S   YESTERDAY'S   YESTERDAY'S
                OPEN       CLOSE        CHANGE

Lucent         100 13/16  101 1/16     + 3 13/16
Ascend         73 5/8     74 7/8       + 2 3/8
AT&T           87 1/2     88 3/16      + 2 3/8
Alcatel        21 7/8     22           + 13/16
Ericsson       26 1/4     26 3/8       + 1 5/16
Motorola       67 1/2     67 1/4       + 1 13/16
DJIA           9367.32    9363.46      + 186.15
NASDAQ         2375.99    2405.55      + 96.05

*******************************************************

*** NEWS IN A NUTSHELL ***      *** LUCENT HERITAGE ***

* New software tool             On Feb. 17, 1998, Lucent
* America's most admired        announced that it would
* Switch lands in winter games  acquire Hewlett-Packard's
* Students visit Bell Labs      local multipoint distribution
* World of Science Seminars     service wireless business
* Client feedback survey        and launch a new Wireless
                                Broadband Networks Division.

****************  LUCENT IN THE NEWS  *****************

STUDENTS VISIT BELL LABS -- Hosted by Lucent Korea,
elementary school students from Korea visited Bell Labs
in New Jersey to explore its advanced science and
technology.  Lucent Korea provided the six-day tour for
the students to encourage their education in science.
[Naeway Economic Daily (Korea), 2/12]
```

Fig. 10. One (or perhaps two) tables embedded in ASCII text.[14] Some general rules, like the use of aligned asterisks or hyphens for rulings, help interpretation of ASCII tables. The frequent (daily?) appearance of such tables, with identical layout but different content, may justify developing specialized algorithms for extracting the information. An important open problem is the detection and isolation of such tables in ASCII text.

[14] From *Lucent Technologies Today*, February 12, 1999.

NEW YORK STOCK EXCHANGE

NYSE INDEXES

NEW YORK (AP) — Closing New York Stock Exchange indexes:

	Close	Chg.
Comp	610.49	—0.19
Indus	761.19	—0.24
Transp	494.71	—5.62
Utility	439.68	+0.75
Finance	549.34	—0.34

WHAT THE NYSE MARKET DID

	Yester-day	Prev. day
Advanced	1,240	1,185
Declined	1,743	1,829
Unchanged	563	568
Total issues	3,546	3,582
New highs	36	58
New lows	96	90

DOW JONES AVERAGES

NEW YORK (AP) — Final Dow Jones averages yesterday:

STOCKS

	Open	High	Low	Last	Chg.
Ind	9902.28	10005.95	9796.99	9890.51	—13.04
Trn	3337.44	3376.11	3242.21	3275.68	—62.80
Uti	303.91	306.48	300.13	303.22	—0.72
Stk	3030.50	3061.77	2985.30	3014.68	—16.16

30 Indus	61,210,600
Tran	8,544,700
Utils	8,781,600
65 Stk	78,536,900

BONDS

	Close	Chg.
DJ AIG Futures	80.34	+1.46
10 Industrials	105.87	—0.30
10 Public Util	102.63	+0.70
20 Bonds	104.25	—0.16

STOCK SALES

Approx final total	663,291,980
Previous day	922,200,000
Week ago	727,270,000
Month ago	718,530,000
Year ago	631,350,000
Two years ago	451,970,000
Year to date	43,374,202,000
To date one year ago	33,969,170,000
To date two years ago	28,938,520,000

BOND SALES

Approx final total	$13,626,000
Previous day	$14,377,000
Week ago	$12,090,000
Month ago	$11,232,000
Year ago	$10,034,000
Two years ago	$22,323,000
Year to date	$759,113,000
To date one year ago	$1,050,662,000
To date two years ago	$1,431,008,000

MOST ACTIVE NYSE STOCKS

NEW YORK (AP) — Sales, closing price and net change of the 15 most active New York Stock Exchange issues trading at more than $1:

Name	Volume	Last	Chg.
AmOnlne s	30,279,300	130	+10 3/4
US Filter	18,371,300	30 3/8	—1/8
Compaq	16,316,100	30 1/8	—5/8
MediaOne	13,143,800	68 1/2	+7 3/4
AT&T	9,387,300	77 3/4	—1 7/8
CHS El	7,415,500	3 15/16	—2 1/16
WarnLm s	7,113,300	66 15/16	—3 3/4
PhilMor	5,984,700	41 11/16	+5/16
IBM	5,948,300	167	—1 9/16
RiteAid	5,777,200	26 3/4	+1 1/8
MicrnT	5,774,300	53	+2 1/2
Lucent	5,254,300	101 1/8	+3/8
CBS	5,094,100	38 5/8	+1 1/16
DataGn	4,661,200	12 5/16	+2 1/8
TycoInt	4,530,500	75 7/16	+3/8

STANDARD & POOR'S

NEW YORK (AP) — Standard and Poor's stock indexes yesterday:

	High	Low	Last	Chg.
S&P 100	653.19	648.44	649.55	—0.56
S&P 500	1303.84	1294.26	1297.01	—2.28
MidCap	363.76	359.82	360.80	—1.51
Indust	1565.34	1552.88	1556.42	—2.67
Transpt	716.73	707.36	708.28	—8.45
Utilities	245.12	243.81	243.99	—0.96
Financial	142.66	141.59	142.22	—0.15
SmallCap	160.66	158.57	158.70	—1.71

Fig. 11. Tables of daily financial results.[15] Some of the quantities are in thousands, others in sixteenths of a dollar. "Industrials" is abbreviated in several ways. The information is condensed and stylized. However, like the previous table, this one can be expected to appear in the same form day after day. Although market information may already available in a completely structured form, like a database, computer queries for other information may require table interpretation.

[15] From *The Trenton Times*, March 23, 1999, pg. D2.

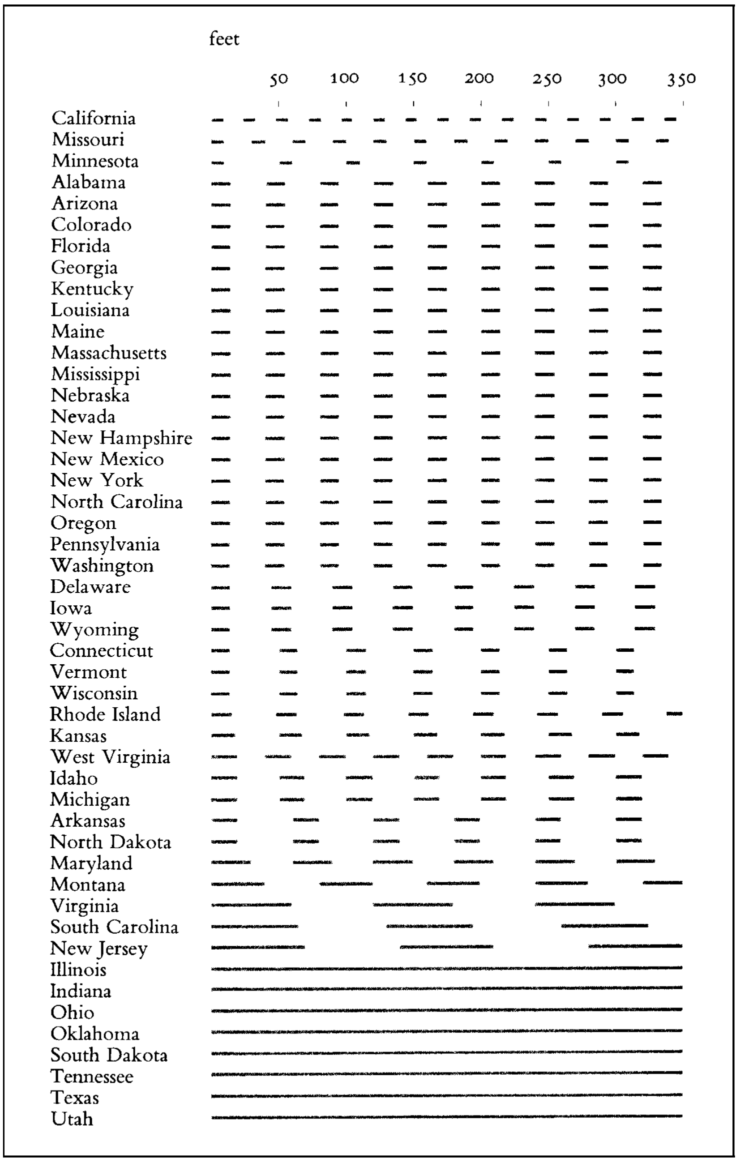


Fig. 12. A table showing standards for painting line stripes on road pavement.[16] This ingenious presentation conveys concisely and visually the length of yellow lane dividers in different states. Automated interpretation is out of the question!

[16] From *The Visual Display of Quantitative Information* by Edward R. Tufte, Graphics Press: Cheshire, CT, 1983, pg. 144.

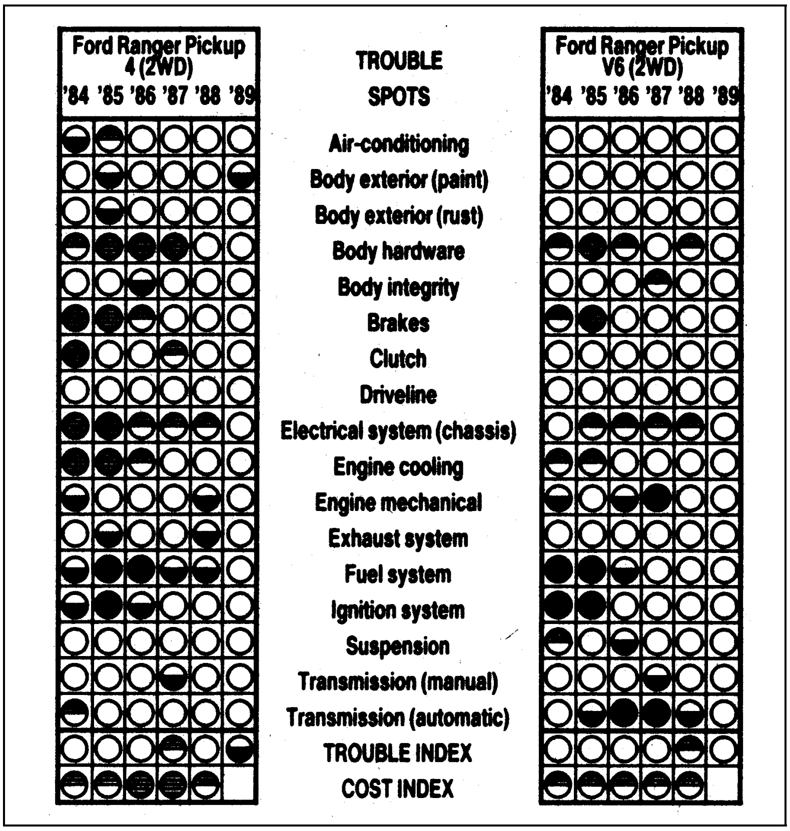


Fig. 13. A table summarizing the reliabilities of two pickup truck models.[17] The use of graphic symbols for cell entries, as in this consumer guide, is not unusual. The legend for the symbols may be far removed from the table itself.

[17] From *Consumer Reports 1991 Buying Guide Issue*, Consumers Union: Mount Vernon, NY, 1990, pg. 159.

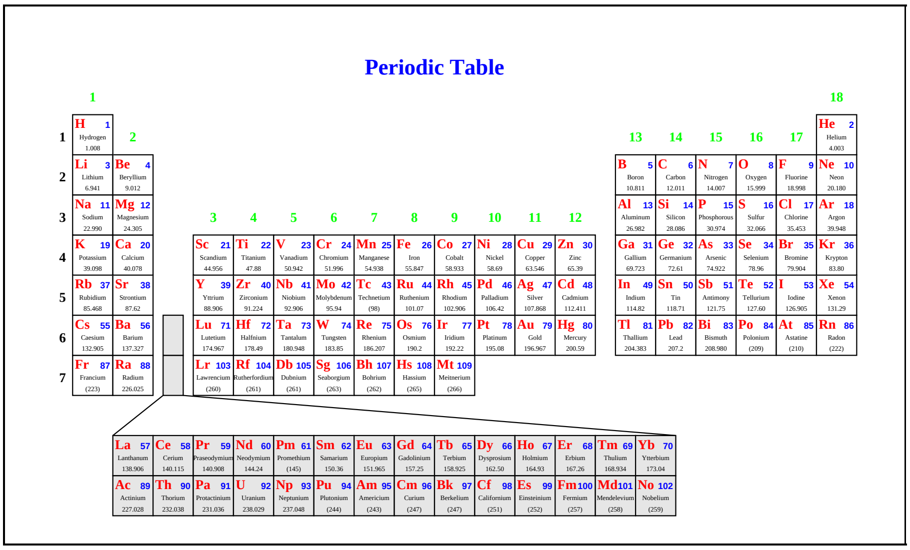


Fig. 14. Periodic Table of the Elements.[18] The Periodic Table is perhaps an extreme example of the challenge that lies ahead for automated table interpretation. It is good to keep in mind that a full understanding of this table may require a lifetime of study.

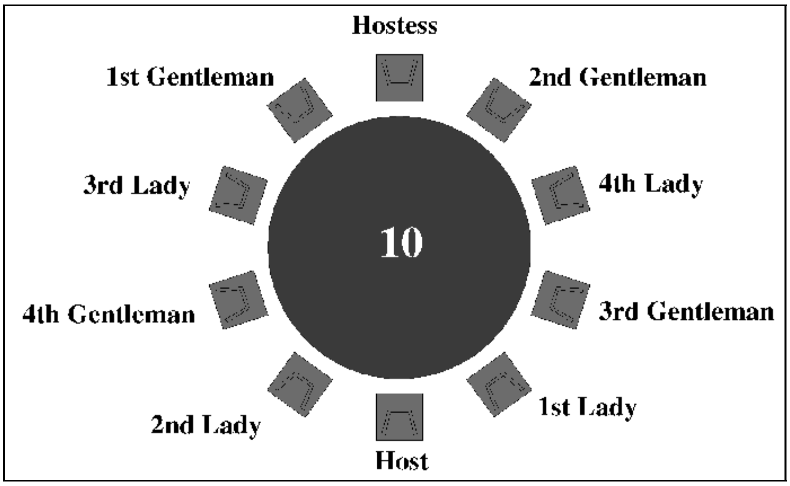


Fig. 15. An example of the wrong kind of "table."[19]

---

[18] From http://www.trends.net/∼mu/misc.html.

[19] From http://www.eglin.af.mil/protocol/tainment/table1.htm.