Text on Tap: the ACL/DCI

Mark Liberman

AT&T Bell Laboratories

Introduction

There has been a recent upsurge of interest in computational studies of large bodies of text. The aim of such studies varies widely, from lexicography and studies of language change to automatic indexing methods and statistical models for improving the performance of speech recognition systems and optical character readers. In general, corpus-based studies are critical for the development of adequate models of linguistic structure and for insights into the nature of language use. However, research workers have been severely hampered by the lack of appropriate materials, and specifically by the lack of a large enough body of text on which published results can be replicated or extended by others.

Recognizing this problem, the *Association for Computational Linguistics* has established the *ACL Data Collection Initiative* (*ACL/DCI*). It provides the aegis of a not-for-profit scientific society to oversee the acquisition and preparation of a large text corpus to be made available for scientific research at cost and without royalties. All materials submitted for inclusion in the collection will remain the exclusive property of the copyright holders (if any) for all other purposes. Each applicant for data from the *ACL/DCI* will be required to sign an agreement not to redistribute the data or make any direct commercial use; however, commercial application of "analytical materials" derived from the text, such as statistical tables or grammar rules, is explicitly permitted. There may be special restrictions on some materials, but only if the restrictions do not compromise the central objective of providing general long-term access for research.

The material in the *ACL/DCI* text corpus will be coded in a standard form based on *SGML*, the *Standard Generalized Markup Language*. Over time, we hope to be able to incorporate annotations reflecting consensually approved linguistic features like part of speech and various aspects of syntactic and perhaps semantic structure. Both the coding and the annotations will be coordinated with the work of the *Text Encoding Initiative* (*TEI*), a project to develop standards for coding and tagging a broad range of different classes of texts to facilitate data interchange and further both research and the language industries. The *TEI* is jointly sponsored by the *ACL*, the *Association for Computers and the Humanities*, and the *Association for Literary and Linguistic Computing*.

Although our initial efforts will concentrate on the collection of American English, we are interacting with groups in other countries with respect to British English and other European languages, and we hope to extend the effort to other language families as well.

History and Current Membership

The ACL/DCI Committee was established in February of 1989. Its current members are Robert Amsler (Bellcore), Bran Boguraev (IBM T.J. Watson Research Center and Cambridge University), Ken Church (AT&T Bell Laboratories), Ed Fox (Virginia Polytechnic Institute & State University), Jim Gallagher (U.S. Department of Justice), Carole Hafner (Northeastern University), Judy Klavans (IBM T.J. Watson Research Center), Mark Liberman (AT&T Bell Laboratories), Mitch Marcus (University of Pennsylvania), Paul Martin (SRI International & MCC), Bob Mercer (IBM T.J. Watson Research Center), Jan Pedersen (Xerox PARC), Paul Roossin (IBM T.J. Watson Research Center), Don Walker (Bellcore), Susan Warwick (ISSCO), and Antonio Zampolli (University of Pisa). Liberman is chairing the committee.

So far, no funding has been obtained or applied for, other than what is implicit in the *pro bono*

efforts of the committee members. Most business has been transacted by email or telephone, thus avoiding travel expenses, and various small out-of-pocket expenses (such as the cost of tapes) have been donated by individual members. However, the problems of acquisition, maintenance and distribution of such a large body of material are beginning to outgrow the bounds of an all-volunteer effort.

Current Status of Collection Efforts

During the eight months since the committee was formed, we have obtained several hundred million words of diverse text. Our current holdings are listed in Appendix A. Perhaps a quarter of this material has been (at least roughly) translated into the SGML-derived format in which it will be distributed.

We have been given the right to distribute (the 1979 edition of) the Collins English Dictionary in electronic form.

Now that we have a fairly large quantity of text, we are beginning to turn our attention to issues of balance in style, genre and topic: what Don Walker calls the "ecology of language." Although what we have is quite diverse, there is obviously a lot left out -- we have no business letters, for instance; no repair manuals; no tabloid newspapers; no movie scripts; no poetry.

Two areas where we plan to concentrate some effort next year are spoken materials, and non-English or multi-lingual text.

Plans for Distribution

We plan to release a "sampler" tape of about 30 million words quite soon. This is as much as will fit on one 12-inch reel of 9-track tape, using easily available (Lempel-Ziv) compression techniques. Of course, we will include the source code to the compression/uncompression programs.

CD-ROM is probably the most appropriate format for distribution of the full database, although 8-mm digital tape also has some fans. We also would like to establish a clearing house for distributing appropriate results of research based on the collection; one promising example is the prospective "tree bank" project at the University of Pennsylvania, which Mitch Marcus is describing at this workshop. Finally, we hope to distribute some simple programs for indexing, word concordancing, manipulating the SGML format, and so forth.

Text Acquisition and Clean-up: Motivations for A Common Effort

In addition to the positive value of having a standard, generally available collection of text (and eventually speech), it's nice to avoid duplicating the often-painful process of obtaining and cleaning text materials. Obtaining the text in the first place may require quite bit of negotiating, once the right person to negotiate with is found. Even if permission is easy to get, actually getting the tapes made and sent may require some pestering and pleading. Once the material arrives, it can sometimes be quite a bit of work to make it usable: the tapes are typically in an undocumented and somewhat complex format, and the text itself may be encoded in an undocumented (and usually proprietary) typesetting language. Also, the correspondance between the logical structure of the text and its typographical structure is often approximate and errorful, so that some intelligence is required if the logical structure is to be recovered.

In Appendix B, I've documented one example where decrypting the donated material was a non-trivial task, namely the Penta dump format of the Library of America volumes. Luckily, most cases are much more straightforward than this; but even in more ordinary examples, the donated text must usually be massaged quite a bit (by programs, of course!).

What is SGML?

SGML stands for Standard Generalized Markup Language. As in the famous joke about the Holy Roman Empire, one may question whether SGML is truly standard, generalized, etc. It is, at least, designated as an international standard by ISO: specifically, ISO 8879. It's designed to allow structural information to be added to a document by embedding user-defined sequences of text characters within the text stream.

"Markup" is the term used to describe codes added to electronically prepared text to define the structure of the text or the format in which it is to appear. "Generalized markup" rises above the details of font definition, type size, exact page layout, etc., to specify more structural concepts such as "heading," "footnote," "emphasis," etc. SGML also makes it possible to define the characters used in a document.

Overall, SGML provides not so much a system for markup as a framework for defining such systems: it is almost infinitely flexible, with the benefits and drawbacks that this entails. The standard itself is far too complex and abstract to permit a brief description -- even its syntax is almost arbitrarily redefinable. It descends towards mortal ken in the form of a **reference concrete syntax** defined in ISO 8879, and one version has acquired the beginnings of a semantics though the publication of proposed tag sets and document type declarations (DTD's) by the Association of American Publishers.

In the examples appended to this paper, the relevent aspects of SGML are mainly its default method for encoding labeled brackets, and its default set of representations for characters outside the ascii (actually ISO 646) set. The start of unit of type *foo* is coded as **<foo>**, and its end (in full form -- we will avoid the thorny issue of *shortrefs*) is **</foo>**. This obviously makes it easy to find foos in a text, at least for a computer, and a simple computer program can transform such a representation into other formats that are easier for humans to read, if desired, as exemplified in samples 3 and 4. Character sequences of the form "*&X;*" (where X is some ascii string not containing ';') are used to encode non-ascii characters. Thus 'e' with an acute accent is "*é*": there are plenty of such examples in the French part of sample 8.

At a minimum, SGML provides a useful interchange format; each researcher can easily write programs to transform such materials into his or her preferred local form. In association with the TEI, we hope to provide a consistent, complete, and well documented tag set, extending the AAP set; this has certainly not yet been done, however.

Appendix A: ACL/DCI Materials, as of 10/89

Department of Energy abstracts
200,00 scientific abstracts, diverse topics (25 million words)

Archives of the Challenger Commission
Transcripts of depositions and hearings about the space shuttle disaster (2.5 million words)

Library of America
American literary classics: 44 volumes (~130 books) promised-- (20 million words)
11 volumes in hand, successfully decrypted:
Twain, Melville, Franklin, Cather, O'Neill, Emerson, Adams, DuBois, etc.

Golden Oldies (public Domain, from various sources):
Tristram Shandy, King James Bible, Federalist Papers etc.

ACL and ACM materials (journals, proceedings etc.)

CSLI (Center for the Study of Language and Information) publications
50-100 reports (8K words each), 5-10 books (80K words each)

Canadian archival materials (Hansard, Supreme Court)
Cleaned-up English Hansard donated by IBM (100 million words)
Original Bilingual Hansard (different time period) obtained directly (200 million words)

U.S. Congressional Record (quantity under negotiation)

Material compiled by Ed Fox at VPISU
U.S. Department of Agriculture Extension Service fact sheets (> 1 million words)

Electronic mail digests: AILIST, IRLIST, etc. (5 million words)
Articles on networking (2 million words)

"Pullum Archive"
About 12K words of administrative policy manuals and 14K words of administrative memos, contributed by Geoff Pullum of U.C.S.C.

Collins English Dictionary (1979 edition)
full text (~3 million words) and various "database" versions

Transcripts of radiologists' reports
Donated by Francis Ganong at Kurzweil AI (about 5 million words)

U.S. Department of Justice JURIS materials

Wall Street Journal
between 25 and 50 million words

Child Language Archive
a diverse collection of transcripts

Negotiations in progress for:

Various on-line technical manuals (e.g. Symbolics)

Other journalistic materials

Boeing repair reports

Other U.S. Government stuff:
(technical manuals, regulations, State Dept. reports...)

Appendix B: Empirical structure of Penta tape format

First level of record structure:

The last 4 of every 514 bytes are two copies of the number of the tape file, in binary: `0 0 0 0` for the first file, `0 1 0 1` for the second file, etc. They must be removed as they cross-cut everything that follows (text, headers, padding, whatever...)

Second level of record structure:

The remaining byte sequence is analyzed as a sequence of chunks like this:
`FF 0 4` : this begins a null-terminated label or filename; it is usually 14 characters long, but not always.
`FB` : this introduces 7 bytes of obscure signification; there is almost always one of these following a null-terminated label.
`FB X Y P Q` : (where X, Y, P, Q are binary bytes of variable value) this introduces X Y bytes of data -- i.e. `FB 1 0 P Q` introduces 256 bytes, `FB 0 C0 P Q` introduces 192 bytes. The `P Q` is variable and of obscure signification. There are 0 or more of these following each filename/FB sequence.
`FC` : following this there are nulls to EOF.

Each labelled chunk is a chapter, a title, a table of contents, etc. The chapters are in not in their normal order in the dump tape; in fact different books may well be intermingled promiscuously; so it is necessary to break out each labelled chunk into a separate file.

Third level of record structure:

Within each labelled chunk (i.e. file), ignore the first 478 bytes. The remainder is divided into records of either 512 or 1024 bytes, depending on whether byte 40 of the record is 0 or 1. In either case, discard the first 46 bytes of the record. If the record is of length 512, byte 41 encodes the useful length of the record as

```
(466 - (239 - X)*2)
```

In other words, for each decrement of byte 41 from 239, ignore two additional bytes at the end of the record. If the record is of length 1024, then there is a first subrecord of 46+500 bytes which is always valid, and then a second subrecord 12+466 bytes. The 12-byte secondary header should be discarded; the amount of valid data in the following 466 bytes is determined by byte 7 in the second subrecord, according to the same formula used above.

Note that end of each 1024- or 512-byte unit is padded, not with nulls, but with a repetition of the corresponding portion of the previous full record.

Last level of record structure.

Within the useful portion of the file, as defined by the above procedure, discard 16 bytes every time `FF FC` occurs. These little cookies seem to function as counters.

The file ends completely when `FF FD` occurs -- stuff following this seems to be random garbage repeating some earlier material.

(Optionally) discard bytes valued 0, 0201, 0202, 0240: they seem to be redundant demarcators of typographical codes.

Results: merely an ordinary unpleasant typographer's tape...

Sample 1: ***Tom Sawyer*****, after removing dump format.**
(newlines are added for display purposes)

```
[j1][ec[dj800][fr[fy66,1][cc22,5,10,14][cj21,24,24][fh1][qc[ar[ep
[ao[bn[fy66,1][cc22,5,10,10][ot5]chapter ii[qc[ot0][ol0][fh1][qr[
ol6.5][ru0][el2.5][xn[ep[ae[bn[fy66,1][cc22,5,10,10][ot5]tom sawy
er[qc[ot0][ol0][fh1][ep[ol6.5][ru0][el2.5][xn[ep[bg[el24][cf1]Lit
erary Classics of the United States [in[el12][qc[ap[ef[j800][j200
]II[ot0][j18][qc[j119]S[cm[cf5]aturday morning[cf1] was come, and
 all the summer world was bright and fresh, and brimming with lif
e. There was a song in every heart; and if the heart was young th
e music issued at the lips. There was cheer in every face and a s
pring in every step. The locust trees were in bloom and the fra~g
rance of the blossoms filled the air. Cardiff Hill, beyond the vi
llage and above it, was green with vegetation, and it lay just fa
r enough away to seem a Delectable Land, dreamy, repose~ful and i
nviting.[epTom appeared on the sidewalk with a bucket of whitewas
h and a long-handled brush. He surveyed the fence, and all gladne
ss left him and a deep melancholy settled down upon his spirit. T
hirty yards of board fence, nine feet high. Life to him seemed ho
llow, and existence but a burden. Sighing, he dipped his brush an
d passed it along the topmost plank; re~peated the operation; did
 it again; compared the insignificant whitewashed streak with the
 far-reaching continent of un-[fjwhitewashed fence, and sat down
on a tree-box discouraged. Jim [cj22,23,24]came skipping out at t
he gate with a tin pail, and sing~ing ''Buffalo Gals.'' Bringing
water from the town pump had always been hateful work in Tom's ey
es, before, but now it did not strike him so. He remembered that
there was company at the pump. White, mulatto and negro boys and
girls were always there waiting their turns, resting, trading pla
ythings, quarreling, fighting, skylarking. And he remembered that
 although the pump [nbwas only a hundred and fifty yards off, Jim
 never got back with a bucket of water under an hour=+=+=m=+=+and
 even then somebody generally had to go after him. Tom said:[ep''
Say, Jim, I'll fetch the water if you'll whitewash some.''[ep
```

Sample 2: *Tom Sawyer*, **SGML Version.**

<ti>TOM SAWYER<ti>
<au>Mark Twain</au>
<chp><no>ii</no>
<s>Saturday morning was come, and all the summer world was bright and fresh, and brimming with life.</s>
<s>There was a song in every heart; and if the heart was young the music issued at the lips.</s>
<s>There was cheer in every face and a spring in every step.</s>
<s>The locust trees were in bloom and the fragrance of the blossoms filled the air.</s>
<s>Cardiff Hill, beyond the village and above it, was green with vegetation, and it lay just far enough away to seem a Delectable Land, dreamy, reposeful and inviting.</s>
<s>Tom appeared on the sidewalk with a bucket of whitewash and a long-handled brush.</s>
<s>He surveyed the fence, and all gladness left him and a deep melancholy settled down upon his spirit.</s>
<s>Thirty yards of board fence, nine feet high.</s>
<s>Life to him seemed hollow, and existence but a burden.</s>
<s>Sighing, he dipped his brush and passed it along the topmost plank; repeated the operation; did it again; compared the insignificant whitewashed streak with the far-reaching continent of unwhitewashed fence, and sat down on a tree-box discouraged.</s>
<s>Jim came skipping out at the gate with a tin pail, and singing <q>Buffalo Gals.</q></s>
<s>Bringing water from the town pump had always been hateful work in Tom's eyes, before, but now it did not strike him so.</s>
<s>He remembered that there was company at the pump.</s>
<s>White, mulatto and negro boys and girls were always there waiting their turns, resting, trading playthings, quarreling, fighting, skylarking.</s>
<s>And he remembered that although the pump was only a hundred and fifty yards off, Jim never got back with a bucket of water under an hour--and even then somebody generally had to go after him.</s>
<s>Tom said:</s>
<s><q>Say, Jim, I'll fetch the water if you'll whitewash some.</q></s>

Sample 3: USDA Fact Sheets, SGML Version

<doc>
<h>Silverfish and Earwigs</h>
<s>Silverfish and earwigs are completely unrelated insects; however, they are two pests which are frequently found in houses.</s>
<s>The silverfish is an insect which grows continually throughout its life and feeds primarily on the glue used in book bindings, cardboard boxes and the like.</s>
<s>So, the places you're likely to have problems with silverfish are around books and bookshelves.</s>
<s>Silverfish are capable of destroying books and other valuable papers so if you find them in your house, you'll need to control them.</s>
<s>Any of the so-called household insecticides are capable of killing silverfish.</s>
<s>The earwig is a different problem.</s>
<s>It's usually a pest in newer subdivisions where houses have been built in what was a wooded area.</s>
<s>Actually, it's not a case of the earwigs moving in on people, but people moving in on the earwigs.</s>
<s>Earwigs don't harm people.</s>
<s>They don't bite, sting, or pinch.</s>
<s>But, they do look bad and they have an unpleasant odor when you step on them.</s>
<s>Of course, they can be a nuisance when they get inside the house.</s>
<s>They feed exclusively on insects and related animals.</s>
<s>Your first line of defense against earwigs should be physically keeping the insects out of your house by blocking areas around doors, windows and other places where they might enter the house.</s>
<s>If you see that you need to resort to chemicals, spray in and around doorways, around patio areas, and indoors along baseboards.</s>
</doc>
<doc>
<h>Advantages of Breast-feeding Your Baby</h>
<s>The American Pediatric Society recommends breast-feeding for infants because of the advantages breast-feeding offers both mother and baby.</s>
<s>Because the nutrients in breast milk are easily digested and ideally suited to a baby's needs, breast milk alone can provide every nutrient a baby needs for the first six months of life.</s>
<s>Usually, no vitamin or mineral supplements are needed, but sometimes doctors may prescribe vitamin D and iron for some babies, and they may urge mothers to give their breast-fed babies fluoride supplements.</s>
<s>Besides getting the nutrients they need, nursing babies also get other health benefits.</s>
<s>Nursing infants have fewer gastronomical illnesses, such as vomiting and diarrhea, and fewer respiratory illnesses, such as colds and infections.</s>
<s>Breast-fed infants also have fewer allergic reactions than do babies fed with bottles.</s>
<s>Another advantage of breast-feeding is that nursing mothers don't have to buy, prepare and sterilize bottles and formula.</s>
<s>Breast-feeding also helps new mothers return to their pre-pregnancy weights because of the extra calories needed for milk production.</s>
<s>Nursing mothers also report that the time spent breast-feeding their babies is a special time for mother and baby.</s>
<s>This closeness is another advantage of breast-feeding.</s>
<s>Breast-feeding gives babies a good start on life.</s>
<s>It can be a wonderful experience for mother and baby.</s>
</doc>

Silverfish and Earwigs

Silverfish and earwigs are completely unrelated insects; however, they are two pests which are frequently found in houses.

The silverfish is an insect which grows continually throughout its life and feeds primarily on the glue used in book bindings, cardboard boxes and the like. So, the places you're likely to have problems with silverfish are around books and bookshelves. Silverfish are capable of destroying books and other valuable papers so if you find them in your house, you'll need to control them. Any of the so-called household insecticides are capable of killing silverfish.

The earwig is a different problem. It's usually a pest in newer subdivisions where houses have been built in what was a wooded area. Actually, it's not a case of the earwigs moving in on people, but people moving in on the earwigs.

Earwigs don't harm people. They don't bite, sting, or pinch. But, they do look bad and they have an unpleasant odor when you step on them. Of course, they can be a nuisance when they get inside the house. They feed exclusively on insects and related animals.

Your first line of defense against earwigs should be physically keeping the insects out of your house by blocking areas around doors, windows and other places where they might enter the house.

If you see that you need to resort to chemicals, spray in and around doorways, around patio areas, and indoors along baseboards.

Advantages of Breast-feeding Your Baby

The American Pediatric Society recommends breast-feeding for infants because of the advantages breast-feeding offers both mother and baby. Because the nutrients in breast milk are easily digested and ideally suited to a baby's needs, breast milk alone can provide every nutrient a baby needs for the first six months of life. Usually, no vitamin or mineral supplements are needed, but sometimes doctors may prescribe vitamin D and iron for some babies, and they may urge mothers to give their breast-fed babies fluoride supplements.

Besides getting the nutrients they need, nursing babies also get other health benefits. Nursing infants have fewer gastronomical illnesses, such as vomiting and diarrhea, and fewer respiratory illnesses, such as colds and infections. Breast-fed infants also have fewer allergic reactions than do babies fed with bottles.

Another advantage of breast-feeding is that nursing mothers don't have to buy, prepare and sterilize bottles and formula. Breast-feeding also helps new mothers return to their pre-pregnancy weights because of the extra calories needed for milk production.

Nursing mothers also report that the time spent breast-feeding their babies is a special time for mother and baby. This closeness is another advantage of breast-feeding.

Breast-feeding gives babies a good start on life. It can be a wonderful experience for mother and baby.

Sample 6: Examples of headings from USDA Fact Sheets

Watering Summer Plants
February Gardening Suggestions
Ajuga
Boxwood
Crape Myrtle
Gardenias
Hollies
Hydrangea macrophylla - Big Leaf Hydrangea
Juniper
Ligustrum
Liriope
Mahonia
Mondo Grass
Nandina
Palms
Photinia
Pyracantha
Vines
Wax Myrtle
Air Layering
Controlling Powdery Mildew on Shrubs
Espaliering Shrubs
Fertilizing Shrubs
Ground Covers
When to Prune Flowering Shrubs
Planting Trees
Radishes
Substitutions for Granulated Sugar
Freezing Poultry
Freezing Baked Goods
Thawing Frozen Foods Safely
Selecting Containers for Freezing
Sack Lunches - Good Safety Tips
Food Safety Tips for Handling Leftovers
Tips on Storing Commercially Canned Foods
Cooking Without Salt
So You Want to Cut Down on Salt . . .
Too Much Sugar - Here's What To Do
Talking Turkey
Selecting a Water Bath Canner
Selecting Canning Jars
Discoloration of Jar Lids
White Sediment in Canned Foods
Cloudy Canning Liquid
Darkening of Food Near the Top of Jars
Canning with Sweeteners Other Than Sugar
Canning Green Beans
Cloudy Pickle Causes
Using Dried Dill or Dill Seeds in Pickles
Substitutions for Flour
Selecting and Using Canning Jar Lids
Making Syrups for Canning
Feeding a Crowd - Food Safety Tips
Freezing Broccoli
Substitutes for Sugar in Jams and Jellies
Making Uncooked Freezer Strawberry Jam
Making Jam in a Microwave
Freezing Sandwiches
Difference Between Honey and Sugar
Aucuba
Problems with Magnolias
Problems with Pines
Birch Problems
Sycamore Problems
Grayish Green Growth on Tree Trunks
Identifying Tree Insect Problems
Leaf Burn in Trees
Mistletoe in Trees
Protecting Trees During Construction
Pruning Mature Trees
Repairing Tree Injuries
Transplanting Trees
Tree Cavities
Hiring a Tree Service
Trees for Shady Locations
Yellowing and Dropping of Leaves
Buying Trees
Why Trees Die in New Subdivisions
Fertilizing Shade Trees
Removing Tree Stumps
Estimating the Value of Trees and Shrubs
Operating a Chain Saw Safely
Removing Trees
Anticipating Shade Needs
Landscape Views From Inside
Low Maintenance Landscapes
Planters in the Landscape
Wood in the Landscape
Wood Fences
Arbors
Growing Cucumbers
Eggplants
Garlic
Herbs
Lettuce
Okra
Onions
Southern Peas
Peppers
Potatoes
Pumpkins and Winter Squash
Shallots
Vegetable Soybeans
Spinach

Summer Squash
Swiss Chard
Tomatoes
Greens
Why Plants Fail to Set Fruit
Caterpillars--Tent, Webworm, Walnut
Slugs and Snails
Aphids
Scale Insects
How Dangerous are Pesticides?
Clothes Moths and Carpet Beetles
Pantry Pests
Silverfish and Earwigs
Advantages of Breast-feeding Your Baby
Contraceptives And Nutrition
Diet for Expectant Mothers
Feeding Solid Foods to Babies
Food Additives
Food Fads and Fallacies
Foods High in Dietary Fiber
Foods on a Low Sodium Diet
Preparing Your Own Baby Food
Risks in Feeding Infants Whole or Skim Milk
Storing Baby Foods
Teenage Pregnancy Nutritional Risks
How Long Does Frozen Food Keep?
Picnic Foods and Safety
Preventing Botulism
Reheat Gravies, Dressings and Meats to Boiling
Defrosting Your Freezer
Defrosting Your Refrigerator
Removing Freezer Odors
Removing Odors from a Refrigerator
Drying Figs
Drying Peaches
Drying Pears
Cooking Dry Soybeans
Cooking Green Soybeans
Cooking Green Vegetables
Cooking Red or Purple Vegetables
Cooking Sweet Potatoes
Making Homemade Ice Cream
Poaching Eggs
Toasting Pecans
Using Leftover Turkey
Cooking Chicken in the Microwave
Converting Recipes for Microwave
Microwave Utensils
Adolescent Weight Control
Calories Do Count
Changing Food Habits
Cooking Methods to Cut Calories
Dangers of Fad Diets
Dieting During the Holiday Season
Exercise and Weight Control
Foods Lower in Calories
Menu Planning for the Dieter
Tricks to Make You Slim
Your Daily Calorie Requirement
Ballpoint Ink
Removing Candle Wax from Fabrics and Surfaces
Removing Chewing Gum From Fabric
Removing Coffee and Tea Stains
Cosmetic Stains
Removing Crayon Stains
Removing Food Stains from Carpet
Removing Food Stains From Upholstered Furniture
Gray Dingies in Your Laundry
Removing Grease from Fabric
Removing Mildew from Clothes
Removing Nail Polish from Fabric and Carpet
Removing Perspiration Stains and Odors from Fabric
Removing Rust Stains
Removing Smoke Odors and Stains from Fabrics
Stain Removal from Carpets and Upholstery
Removing Stains from Clothing
Removing Urine Stains from Carpets and Upholstery
Removing Urine Stains from Clothing and Linens
Auto Insurance--Buying
Auto Insurance--How Much Do You Need?
Coupons Save Money
Stretch Your Dollars
Door-to-Door Sales
Enjoying Your Income
Family Records--What to Keep and For How Long
Family Records--Storage
Health Insurance--Buying
Homeowners Insurance: How Much do You Need?
How to Buy Appliances
Ordering Merchandise by Mail
Planned Spending Buys More
Safe Debt Load
Shopping for Credit
Teaching Children to Use Money
Women and Credit
Supermarket Scanners
Controlling Your Finances While Unemployed
Flood Insurance
Conserving Water In The Home
Water Filters
Water Hardness
Water Softeners
Laundry - Saving Energy
What is 4-H?
Virginia Extension Homemakers Council
Reclaiming Old Shrubs and Small Trees

Sample 7: Canadian Hansard Material, SGML Version.

```
<timestamp id=canpar/860417.E>
<CAPS>canadian charter of rights and freedoms</CAPS>
</SC>Fourth Anniversary of Proclamation</SC>
<speaker>Hon. Benont Bouchard (Secretary of State of Canada)</speaker>

<p><s>Mr. Speaker, I would like to bring to the attention of the House that
today, as Hon. Members are no doubt aware, we are celebrating the
anniversary of the proclamation of the Canadian Charter of Rights and
Freedoms which took place on April 17, 1982, and also of the coming
into effect a year ago of the provisions guaranteeing equality for all
members of our society.</s></p>

<p><s>It is a day on which Hon. Members will come together to commemorate a
commitment to equality, social justice, tolerance and fairness for all
Canadians in keeping with basic standards of human rights and
fundamental freedoms.</s></p>

 ...
</timestamp id=canpar_860417.E>

<timestamp id=canpar_860417.F>
<CAPS>LA CHARTE CANADIENNE DES DROITS ET LIBERT&Eacute;S</CAPS>
<SC>Quatri&egrave;me anniversaire de la proclamation</SC>
<speaker>L'hon. Benont Bouchard (secr&eacute;taire d'&Eacute;tat du Canada)</speaker>

<p><s>Monsieur le Pr&eacute;sident, je voudrais porter &agrave; l'attention
de la Chambre que nous c&eacute;l&eacute;brons aujourd'hui, comme le
savent les honorables d&eacute;put&eacute;s, l'anniversaire de la
proclamation de la Charte canadienne des droits et libert&eacute;s qui
a eu lieu le 17 avril 1982, ainsi que son parach&egrave;vement, il y a
un an, avec l'entr&eactue;e en vigueur des dispositions garantissant
l'&eacute;galit&eacute; &agrave; tous les membres de notre
soci&eacute;t&eacute;.</s></p>

<p><s>Aujourd'hui, les d&eacute;put&eacute;s rappellent l'engagement que nous
avons pris d'assurer &agrave; tous les Canadiens
l'&eacute;galit&eacute;, la justice sociale, la tol&eacute;rance et
l'&eacute;quit&eacute;, en conformit&eacute; des normes admises en
mati&egrave;re de droits de la personne et de libert&eacute;s
fondamentales.</s></p>
 ...
</timestamp id=canpar_860417.F>
```

Sample 8: Emerson, Formatted Version.

Nature

Introduction

Our age is retrospective. It builds the sepulchres of the fathers. It writes biographies, histories, and criticism. The foregoing generations beheld God and nature face to face; we, through their eyes. Why should not we also enjoy an original relation to the universe? Why should not we have a poetry and philosophy of insight and not of tradition, and a religion by revelation to us, and not the history of theirs? Embosomed for a season in nature, whose floods of life stream around and through us, and invite us by the powers they supply, to action proportioned to nature, why should we grope among the dry bones of the past, or put the living generation into masquerade out of its faded wardrobe? The sun shines to-day also. There is more wool and flax in the fields. There are new lands, new men, new thoughts. Let us demand our own works and laws and worship.

Undoubtedly we have no questions to ask which are unanswerable. We must trust the perfection of the creation so far, as to believe that whatever curiosity the order of things has awakened in our minds, the order of things can satisfy. Every man's condition is a solution in hieroglyphic to those inquiries he would put. He acts it as life, before he apprehends it as truth. In like manner, nature is already, in its forms and tendencies, describing its own design. Let us interrogate the great apparition, that shines so peacefully around us. Let us inquire, to what end is nature?

All science has one aim, namely, to find a theory of nature. We have theories of races and of functions, but scarcely yet a remote approach to an idea of creation. We are now so far from the road to truth, that religious teachers dispute and hate each other, and speculative men are esteemed unsound and frivolous. But to a sound judgment, the most abstract truth is the most practical. Whenever a true theory appears, it will be its own evidence. Its test is, that it will explain all phenomena. Now many are thought not only unexplained but inexplicable; as language, sleep, madness, dreams, beasts, sex.

Philosophically considered, the universe is composed of Nature and the Soul. Strictly speaking, therefore, all that is separate from us, all which Philosophy distinguishes as the not me, that is, both nature and art, all other men and my own body, must be ranked under this name, Nature. In enumerating the values of nature and casting up their sum, I shall use the word in both senses;--in its common and in its philosophical import. In inquiries so general as our present one, the inaccuracy is not material; no confusion of thought will occur. *Nature*, in the common sense, refers to essences unchanged by man; space, the air, the river, the leaf. *Art* is applied to the mixture of his will with the same things, as in a house, a canal, a statue, a picture. But his operations taken together are so insignificant, a little chipping, baking, patching, and washing, that in an impression so grand as that of the world on the human mind, they do not vary the result.

The workshop was held to collect current data on the experience with primary water stress corrosion cracking (PWSCC) of steam generator tubing and the related laboratory investigations. Thirty-two presentations were given covering field experience, correlations of laboratory data on the field, and relationship of material microstructure, stress, and environment to PWSCC. The emphasis of the workshop was more on the fundamentals associated with PWSCC yet culminated with several presentations on remedial measures.

The /sup 252/Cf neutron spectrum has been measured at high energies with a miniature ionization chamber and two different NE213 neutron detectors. The {gamma}-ray background and the main cosmic background were suppressed by applying an efficient pulse shape n/({gamma},{mu}) discrimination. On the basis of the two-dimensional spectroscopy of neutron time-of-flight and scintillation pulse height, the sliding bias method has been used to minimize experimental uncertainties. The experimental data corrected for several systematic influences confirm earlier results which show a trend similar to the NBS evaluation. However, the final spectra obtained for both neutron detectors exhibit negative deviations (up to -10%) from the NBS curve in the 6-12 MeV range. Finally, the experimental results of this work are compared with various statistical-model approaches to the /sup 252/Cf(sf) neutron spectrum. 16 refs, 16 figs, 3 tabs.

The effects of ion-implantation on the surface mechanical properties of ceramics is investigated. Changes in hardness and indentation fracture that occur in reaction-bonded silicon carbide, sialon, partially stabilized zirconia and WC are all described. These modifications are correlated to the structural changes brought about by the implantation process.

A vertical fracture type reservoir is assumed with a production well, for studying the effect of heat conduction to the response curve of the tracer using a mathematical model. With the inlet and outlet of the model located on the center axis of the fracture, flow in the fracture is always stationary. The tracer moves along the flow while dispersing mixedly. Inflow rate is 4 kg/s with an assumed inpermeable temperature boundary rock of 1 (5 m) thickness around the fracture. Tracer responses are calculated with distributed temperatures and even temperature in the fracture. The width of flow inlet does not influence tracer response. However, the height of the flow inlet and temperature distribution in the fracture greatly affect the response. (18 refs, 9 figs, 2 tabs)

Tools, equipment and weapons contaminated with radioactive, toxin, biological and/or chemical contaminants are deposited in a cleaning chamber and are sprayed with a solvent under high pressure. The solvent dislodges particulate contaminants and dissolves chemical agent contaminants and the solvent so sprayed containing both suspended and dissolved contaminants is drained to a distillation means. Within the distillation means there is a neutralizing agent which deactivates the biological and toxin contaminants and chemically oxidizes the chemical contaminants removed from the item being decontaminated in the cleaning chamber. Pure solvent vapor generated in the distillation means is condensed to a solvent tank for reuse in the spraying operation for further decontamination. Drying of the tool, equipment or weapon being decontaminated is accomplished by circulating hot, unsaturated solvent vapor through the cleaning chamber and about the item being decontaminated.

We have studied the growth and metabolism of Syntrophomonas wolfei in pure culture with crotonate as the energy source. S. wolfei grows in crotonate mineral salts medium without rumen fluid with cobalamin, thymine, lipoic acid and biotin added. However, after four to six transfers in this medium, growth ceases, indicating that another vitamin is required. The chemically defined medium allows large batches of S. wolfei to be grown for enzyme purification. All the enzymes involved in the oxidation of crotonyl-CoA to acetate have been detected. The pure culture of S. wolfei or coculture of S. wolfei grown with crotonate contain high activities of a crotonate: acetyl-CoA CoA-transferase activity. This activity is not detected in cocultures grown with butyrate...

Sample 9: Acronym Dictionary Derived from DOE Abstracts.

1	A-T	ataxia telangiectasia
2	AA	Anti-proton Accumulator
1	AA	Arachidonic acid
1	AA	Automobile Association
1	AA	additional absorption
1	AA	allyl alcohol
1	AA	amino acid
1	AA	andesine anorthosite
6	AA	arachidonic acid
1	AA	aristolochic acid
2	AA	ascorbic acid
1	AA	atomic absorption
1	AA	atomic adsorption
1	AA	azelaic acid [HOOC(CH/sub 2/)/sub 7/COOH]
1	AAA	abdominal aortic aneurysm
1	AAA	amino acid analogs
1	AAAC	all-aluminium alloy conductor
1	AAB	Additional absorption bands
1	AABW	Antarctic Bottom Water
2	AACE	Automation and Control Experiment
1	AACS	Airborne Activity Confinement System
1	AACS	Automated Access Control System
1	AACSR	all-aluminium conductor steel reinforced
2	AAD	atlanto-axial dislocation
2	AAEC	Australian Atomic Energy Commission
1	AAFES	Army-Air Force Exchange Service
1	AAGs	accumulations of autoradiographic grains
1	AAH	Accident Analysis Handbook
1	AAI	aristolochic acid I
1	AAII	aristolochic acid I (AAI) and II
1	AALA	American Association for Laboratory Accreditation
1	AAM	atmosphere angular momentum function
1	AAOE	Airborne Antarctic Ozone Experiment
1	AAOO	American Academy of Ophthalmology and Otolaryngology
1	AAP	alumina-aluminum phosphate
1	AAPB	American Association of Pathologists and Bacteriologists
1	AAPM	American Association of Physicists in Medicine
1	AAQS	ambient air quality standards
1	AAR	Association of American Railroads
1	AAR	alkali aggregate reaction
1	AAS	Andersen air sampler
6	AAS	atomic absorption spectrometry
2	AAS	atomic absorption spectrophotometry
3	AAS	atomic absorption spectroscopy
2	AASB	aerodynamically air staged burner
1	AAT	asymptomatic autoimmune thyroiditis
1	AATR	Automatic Alarm Testing Robot
2	AAV	adeno-associated virus
1	AAV	assault amphibious vehicle
1	AB	Agua Boa
1	AB	Aharonov-Bohm
1	AB	asbestos bodies
1	AB	atomic bremsstrahlung
1	ABA	abscisic acid
1	ABAE	Adult bovine aortic endothelial
1	ABB	Allender, Bray and Bardeen
2	ABB	Asea Brown Boveri
1	ABC	Analysis of Benefits and Costs
1	ABCDHW	Ames-Bologna-CERN-Dortmund-Heidelberg-Warsaw
1	ABE	acetone-butanol-ethanol
1	ABG	asymptotic branch of giants
1	ABL	atmospheric boundary layer
1	ABLS	alpine breaker liner support

MR. MOLESWORTH: This is Investigator John R. Molesworth, that's M-o-l-e-s-w-o-r-t-h, interviewing Mr. Kapp at the offices of Morton-Thiokol.

Mr. Kapp, will you give your name and position here and a brief description of your duties.

MR. KAPP: That's Jack Kapp --

MR. MOLESWORTH: Spell your name, also.

MR. KAPP: K-a-p-p, "P" as in Paul. My position with Morton-Thiokol is Manager of the Applied Mechanics Department.

I am responsible in general for all of the structural analysis of all of the complements, rocket motor complements, that Thiokol develops. I have approximately 100 people working in my department.

MR. MOLESWORTH: How long have you worked here?

MR. KAPP: I've been at Thiokol about 27-1/2 years.

MR. MOLESWORTH: And you worked on the solid rocket motor?

MR. KAPP: I've been in this general area of work, structural analysis, for the full 27-1/2 years. Prior to that time, I have a bachelor's degree in mechanical engineering and a master's degree in mechanical engineering that I obtained in 1969. I have been in this line of work for the whole 27-1/2 years, functioning as lead engineer, unit chief, then section supervisor, and then finally department manager.

As far as the Shuttle Program specifically is concerned, I have been involved with it from the very inception, being assigned to the proposal team and working in various capacities with the structural analysis of the rocket motor case and more lately the nozzle and the propellant since that time.

MR. MOLESWORTH: Do you recall the telecon which was held on January 27th concerning the launching of 51-L?

MR. KAPP: Yes, sir, I do. I did not take any personal notes at that time but I well remember the conversation.

MR. MOLESWORTH: Directing your attention to January 27th, when was the -- when did you first learn of any concern over the temperature at Cape Kennedy, and how were you involved in resolving or researching any problem concerning the temperature?

MR. KAPP: Okay, let me retrogress just a little bit and lay some background.

I have two individuals that work for me, one by the name of Roger Boisjoly, B-o-i-s-j-o-l-y. Mr. Boisjoly is on my staff, and for the last many months has been given the specific assignment of working very closely with the SRM seals.

I also have another individual, Mr Arnold Thompson, who is a section supervisor under me who is directly responsible for the structural analysis and design of the case.

These two individuals have been very involved in the seal work for about six months to a year, even prior to that, but intensively for the last year.

Because of their experience level, both of which are in the area of 30 years, I have not personally been as vitally involved in that area as I might if a more junior engineer had been involved.

Now, having said that, about 3:45 on the afternoon of the 27th of January, Mr. Arnold Thompson, one of my section supervisors, come into my office and indicated to me that they had been meeting most of the afternoon and that they had some very serious concerns about the temperatures that were predicted at the Cape, they being primarily the O-ring task force that had been set up to investigate this problem and develop solutions.

He give me about a 15-minute synopsis of what his concerns were, indicating that the decision had been made to take those concerns to higher management. At the time I told --

MR. MOLESWORTH: Do you recall who he was meeting with, the names of the individuals?

MR. KAPP: I know that -- I know that Brian Russell was involved. I know that Don Ketner was involved, that Roger Boisjoly was involved, and there may have been others.

MR. MOLESWORTH: Okay, and was Arnie Thompson the spokesman for the group, or --

MR. KAPP: I think not. Arnie just -- I had to be informed, and I think Arnie just decided the time has come to involve me.