Syllabus

Calendar Overview by part of course

Read only if you need another look at the overall structure of the course

Calendar Overview by part of course

Note: This part **is repetitive with the calendar**; it gives a different arrangement which some students might find helpful.

Part 1. Foundations. Knowledge and knowledge representation (Lectures 1.1 - 2.2)

The course starts out with the **fundamentals**: Combining ideas from database management, artificial intelligence, and cognitive psychology, we explore the nature of knowledge and its structure and representation in information systems and in the mind. Out of this exploration evolve general principles that apply to any kind of information/knowledge.

These ideas are **applied** and made more concrete

- in three assignments that illustrate the application of information structure to searching;
- throughout the course.

Lectures		
1.1	Jan. 19	Introduction and overview. Information professionals in the 21st century
1.2		Information systems and information structure
2.1	Jan. 26	The nature of knowledge
2.2		Knowledge representation

	Assignment	Assigned	Due
These assignments guide you in the exploration of three information systems, illustrating how information structure is used for navigation and query-based search.			
1	Hypermedia exploration: Perseus and Freebase (2.5 hours)	Jan. 19	Feb. 2
2	Bibliographic retrieval system exploration: MEDLINE (3 hours)	Jan. 19	Feb. 2
3	Online catalog search exercise (1.5 hours)	Jan. 19	Feb. 2

Part 2. Information retrieval: General principles and methods (Lectures 3.1 - 5.1, also 11.2)

This part begins with an **overview of the structure of information systems**, systems that bring information or knowledge to the people or organizations or computer systems who need it to solve problems. This is followed by a discussion of information system objectives. The result is an overall framework for the discussion of individual information system functions and components not just in this course but in other courses as well.

Building on the conceptual foundation of Part 1, Part 2 then introduces a **general information structure model** that provides an integrated view of different approaches to information retrieval (IR). It then discusses **data schemas and formats** and the data structure and search component of IR systems, all on a general level, laying out principles to be applied to specific types of systems later in the course. The later lecture 11.2, Indexing and system performance, rounds out this part.

Lectures		
3.1	Feb. 2	The structure of information systems.
3.2		Objectives and performance measures for information systems
4.1	Feb. 9	An integrated information structure model
4.2		Data schemas and formats
5.1	Feb. 16	Access to information: data structure & search modes. Retrieval as prediction. Ranking
11.2	Apr. 7	Indexing and system performance

	Assignment	Assigned	Due
4	Analytical description of an information system	Feb. 2	Feb. 9
5	Developing a conceptual data schema	Feb. 9	Feb. 16
6	Restructuring a semantic network	Feb. 16	Mar. 2
	Short description of term paper	Feb. 16	Mar. 2

Part 3. The nature, design, and management of documents and records (Lect. 5.2 - 7.2)

Part 3 applies the general principles to the specific case of documents and records, from plain text on paper to multimedia Web sites. It explores **how knowledge**, a complex web of interrelationships among entities, **is (re)presented in text and images**; put differently, it explores the structure of documents and principles of document design for improved communication. It examines how text/document structure affects assimilation and understanding. It covers text types; text analysis, including **natural language processing** (specifically syntactic and semantic parsing) and data extraction; application of frames to the analysis of document macro structure; **document design for people**, expressing the internal conceptual structure through external form; and briefly mentions markup languages to make documents understandable for computers.

Part 3 then uses general information structure principles and insights into the nature of documents to elucidate the problems of **describing/cataloging documents** and designing library and Web catalogs – the problem of metadata, with a look to supporting users most effectively.

Lectures		
5.2	Feb. 16	**Document function, structure, analysis, and design.** 5.2A Knowledge (re)presentation in text and images. Text linguistics.
6.1 - 6.2	Feb. 23	**Micro** 5.2BText analysis. 6.1A Natural language processing, syntactic and semantic parsing **Macro** 6.1B Document macrostructure. 6.2A Document design. 6.2B Markup languages
7.1 - 7.2	Mar. 2	**Bibliographic and record control.** General issues, description, entries and access

	Assignment	Assigned	Due
7	Applying linguistic techniques to retrieval problems	Feb. 23	Oct 12
8	Problems of entry	Mar. 2	Mar. 23
9	Descriptive cataloging of four documents	Mar. 2	Mar. 23+
10	Indexing of three documents and prep for Lecture 8.1 (Belongs to Part 4)	Mar. 2	Mar. 8-10

Part 4. Classification and subject access (Lectures 8.1 - 14.2)

While Parts 1 - 2 deal with the access to information and documents from all kinds of access points and Part 3 focuses on formal or "descriptive"access points, Part 4 focuses on subject access. It applies the principles of information structure and user orientation to an **analysis of knowledge organization systems (KOS)** – classification schemes and thesauri, taxonomies, ontologies. Part 4 relies mainly on assignments designed to help you explore such schemes to (1) reinforce understanding of the general principles and teach the skill of analyzing such schemes and (2) help you get familiar with a few widely used schemes.

Lectures.		
8.1	Mar. 8 - 10	Small Groups 1. Explorations in subject access. (to be scheduled)
8.2		Vocabulary control. Lexical relationships. Index language functions
9.1	Mar. 22 - 24	Small Groups 2. Index language structure 1: conceptual (to be scheduled)
9.2		Application of index language structure to searching
10.1	Mar. 29 - 31	Small Groups 3. On constructing a hierarchy from facet combination (to be scheduled) Index language structure 2: database organization
10.2		Brief discussion of Assignment 13: Subject cataloging and searching practice
11.1	Apr. 7	Index language structure 2. Database organization
11.2		Indexing and system performance
12.1-12.2	Apr. 14	Introductory discussion and in-class exercise: DDC. Short Media Streams Demo
13.1-13.2	Apr. 21	Introductory discussion and in-class exercise: Yahoo, LCC, and LCSH
14.1	Nov 30	Exploration of classification schemes and thesauri
14.2		Concluding discussion and comparison of classification schemes and thesauri

	Assignment	Assigned	Due
10	Indexing of three documents and preparation for Lecture 8.1	Mar. 2	Mar. 8 - 10
11	Request-oriented indexing	Mar. 9	Mar. 23
***	Take-home midterm exam	Mar. 9	Mar. 23
12	Conceptual analysis and synthesis	Mar. 23	Mar. 29 - 31
13.1	Dewey Decimal Classification (DDC)	Apr. 7	Apr. 14
13.2	ERIC Thesaurus	Apr. 14	Apr. 21
13.3	Library of Congress/Sears Subject Headings (LCSH)	Apr. 14	Apr. 21

13.4	Yahoo: Yahoo (or DMOZ) classification OR LCC: Library of Congress Classification OR MediaStreams iconic classification OR Own choice	Apr. 21	May 4

Lecture 1.1 Supplement

Phone: 202.421.7609
Email: allison.denny@gmail.com

Allison Denny

Experience

DataStream Content Solutions 9/2006-Present
Senior process engineer

- Project lead: development of an XML-based electronic publishing system for the Office of the Legisla-tive Counsel of the U.S. House of Representatives. Analyze new and legacy legislative data and define data model requirements. Map content into data models; create, maintain and document DTDs; define conversion specifications to XML and print formats. Gather and document business and technical re-quirements and support editorial workflow processes. Manage software release schedules and testing.

Discovery Communications, Inc. 3/2006-8/2006
XSLT subcontractor

- Wrote XSLT stylesheets for converting SQL data into wordprocessingML. Defined input XML schema and XML data targets, documented data transforms.

Education Resources Information Center, U.S. Department of Education 7/2004-3/2005
Lexicography subcontractor

· Facilitated the implementation of automatic indexing software and a revised workflow process.

LexisNexis 1/2001-6/2004
XML data architect

- Responsible for consistency and traceability of XML data elements through the development of a data-driven publication management system. Built and maintained an environment of related XML data models, including DTDs, UML diagrams, XSLT stylesheets, templates and technical specifications. Acted as project lead for data-intensive software release cycles. Supported users defining data requirements and business processes. Wrote documentation and training materials.

National Public Radio 1/1998-8/1999
Broadcast librarian

· Catalogued and indexed daily NPR programming. Provided reference service using NPR database.

Education

Georgetown University 9/2003-12/2007

Communication, Culture, and Technology

- Master of Arts. Interdisciplinary program exploring media and technology from social, economic, political, and cultural perspectives. Seminars included Knowledge Management, Computer-Mediated Communication, Technologies of the Text.

University of Maryland, College Park 9/1999-12/2000

College of Information Studies

- Master of Library Science. Concentration in Information Organization with coursework including Information Structure, Construction of Thesauri, Abstracting and Indexing, Database Design.

Northwestern University 9/1991-6/1995

- Bachelor of Arts, honors. Major in American Studies, minor in French.

Technology

- XML-related technologies: XML, XPath, XSLT, DTD, XSD, RELAX NG, schematron, CSS, UML; familiar with RDF/OWL, topic maps, DITA.
- XML-related tools: XMetaL, XML Spy.

Tamar Donovan

6058 SUNNY SPRING • COLUMBIA MD 21044
PHONE 410-772-9049 • E-MAIL KORODON@AOL.COM

WORK EXPERIENCE

05/2005 –

Independent contractor /consultant

- Currently: Metadata Librarian/Information Architecture Consultant, The Electronic Scriptorium Ltd.

Some projects I have participated in:

- Organizing a taxonomy of job titles for a job-search website;
- Designing metadata structure for the digital repository of a symphony orchestra;
- Designing a database to support construction of a names memorial;
- Designing metadata structure for the digital repository of an academic journal.

08/2004 – 05/2006 Gibson Library, J.H.U. Applied Physics Laboratory. Laurel, Maryland

Circulation Clerk (part-time)

- Circulation; technical services; reference and research assistance; special projects.

American Embassy, Tashkent. Tashkent, Uzbekistan. *Consular Associate*
2001 - 2002

· Conducted non-immigrant visa interviews; adjudicated visas; assisted in anti-fraud investigations.

American Consulate, Vladivostok Vladivostok, Russia
1998 - 2000
Consular Associate

· Conducted non-immigrant visa interviews; adjudicated visas; assisted in anti-fraud investigations.

Ankara Community Library. Ankara, Turkey. *Volunteer*
1996-1998

· Performed circulation and shelving duties.

Indiana University Library System Bloomington, Indiana
1987 - 1988
Assistant to Modern Languages Librarian

EDUCATION

2003 - 2005 University of Maryland, College of Library and Information Science. *M.L.S., May, 2005.* Track: Information Access and Use.

Indiana University, Department of Central Eurasian Studies. Major fields of study:
1986 – 1988
Hungarian language and literature, comparative studies of Uralic and Altaic language and peoples.

The Johns Hopkins University, B.A., Classics.
1984 - 1986
Inducted into Phi Beta Kappa, May, 1986. Graduated with Departmental and university honors.

LANGUAGES

Fluent speaker of Russian. Reading knowledge of several European languages.

Salaries of reporting professionals* by area of job assignment
Library Journal Oct. 2008, 2007 numbers. Full-time placements

ASSIGNMENT	No.	% of Total	Low Salary	High Salary	Average Salary	Median Salary
Acquisitions	18	1.3%	26K	70K	42K	39K
Administration	62	4.6%	18K	121K	44K	39K
Adult Services	44	3.3%	19K	48K	36K	36K
Archives	59	4.4%	14K	65K	40K	40K
Automation/Systems	21	1.6%	30K	93K	52K	48K
Cataloging & Classification	76	**5.6%**	18K	70K	40K	40K

ASSIGNMENT	No.	% of Total	Low Salary	High Salary	Average Salary	Median Salary
Children's Services	75	5.5%	20K	55K	38K	38K
Circulation	51	3.8%	19K	55K	32K	33K
Collection Development	18	1.3%	30K	53K	41K	41K
Database Management	10	0.7%	24K	75K	41K	36K
Electronic or Digital Services	51	3.8%	24K	70K	45K	43K
Government Documents	8	0.6%	32K	50K	39K	38K
Indexing/Abstracting	6	0.4%	26K	26K	26K	26K
Info Technology	44	3.3%	32K	150K	53K	47K
Instruction	41	3.0%	17K	70K	42K	41K
Interlibrary Loans/ Doc. Del.	19	1.4%	21K	45K	34K	32K
Knowledge Management	7	0.5%	28K	51K	40K	41K
Other	110	8.1%	15K	115K	46K	43K
Reference/Info Services	293	**21.6%**	19K	70K	41K	40K
School Library Media Spec.	191	**14.1%**	25K	91K	44K	43K
Solo Librarian	51	3.8%	25K	57K	39K	39K
Usability/Usability Testing	15	1.1%	50K	90K	75K	78K
Web Design	1	0.1%	45K	45K	45K	45K
Youth Services	83	**6.1%**	20K	52K	36K	36K
TOTAL	**1354**	**100.00**	**14K**	**150K**	**42K**	**40K**

Library Jobs by Level, ALA survey 2008. Average salary

2008 *ALA-APA Salary Survey: Librarian – Public and Academic* (*Librarian Salary Survey*)

Job title	Public	Academic
Director/Dean/Chief Officer	86K	95K
Deputy/Associative/Assistant Director	73K	80K
Dept Head/Branch Mgr/Coordinator/Senior Mgr	61K	61K
Manager/Supervisor of Support Staff	52K	54K
Librarian Who Does Not Supervise	48K	55K
Beginning Librarian	43K	45K

www.ala-apa.org/salaries/SalarySummary2008.pdf (Tables 1 and 2)

Some jobs in other environments (numbers from **www.payscale.com**, swz.salary.com, cbsalary.com, 2003 compilation by Roberta Shaffer)

Job title	From	To	Source
Chief Knowledge Officer	66K	130K	payscale
Chief Information Officer	90K	153K	payscale
Information Technology (IT) Manager	48K	92K	payscale
Chief Information Security Officer	127K	184K	salary
Information Architect	39K	103K	payscale
Ontologist	62K	84K	payscale
Senior content specialist	53K	74K	salary
Information Analyst	46K	126K	cbsalary
Consumer Information Director	55K	118K	cbsalary
Archivist	38K	52K	payscale
Strategic Information Planner	57K	75K	RS 2003
Business Intelligence Manager	55K	90K	RS 2003
Manager, Campus Technology and Academic Computing	62K	135K	RS 2003

Legal Information Specialist	50K	80K	RS 2003
Sarbanes-Oxley Compliance Manager, IT	89K	97K	payscale

* Note that numbers from payscale.com are the **Median** Salary by Years Experience charts

Supplement for Lecture 1.2

Example 3. **Medline, a bibliographic information system**

Medline, from the National Library of Medicine, is the premier bibliographic system in medicine

Purpose	Find documents on a given subject Answer the question: what Documents X <*dealsWith*> a given Subject?

System for simple subject search

Two types of facts
A title facts **B indexing facts** (index terms)

Question	What documents deal with Hearing tests? Document X <*dealsWith*> Hearing tests
Facts	A1 Document 1 <*hasTitle*> Measurement of acoustic impedance in the ear canal B1 Document 1 <*dealsWith*> Acoustic impedance tests B2 Document 1 <*dealsWith*> Computer simulation B3 Document 1 <*dealsWith*> Hearing--physiology A2 Document 2 <*hasTitle*> Optimization of automated hearing test algorithms B4 Document 2 <*dealsWith*> Algorithms B5 Document 2 <*dealsWith*> Auditory threshold B6 Document 2 <*dealsWith*> Computer simulation B7 Document 2 <*dealsWith*> Hearing tests A3 Document 3 <*hasTitle*> Expert systems for medical diagnosis B8 Document 3 <*dealsWith*> Diagnosis B9 Document 3 <*dealsWith*> Expert systems B10 Document 3 <*dealsWith*> Neural networks (computer) A4 Document 4 <*hasTitle*> New standard enhances efforts in hearing conservation. B11 Document 4 <*dealsWith*> Audiometry B12 Document 4 <*dealsWith*> Data interpretation, statistical B13 Document 4 <*dealsWith*> Ear protective devices--standards B14 Document 4 <*dealsWith*> Hearing loss, noise-induced--prevention and control
Rules	None
Answer	Document 2 (due to fact B7 document 2 <*dealsWith*> Hearing tests)

But things are not so simple. There are actually more documents on the topic, but they deal with specific hearing tests rather than hearing tests in general. To find these documents the system needs additional knowledge, an additional type of facts, namely hierarchical relationships between concepts. Such facts are available in the Medical Subject Headings published by National Library of Medicine. The database Medline uses the inclusive searching method discussed below.

System for more complete subject search exploiting knowledge of hierarchy (fact type C)

Three types of facts
A title facts B indexing facts (index terms) **C concept hierarchy facts**

Question	What documents deal with Hearing tests? Document X <*dealsInclusivelyWith*> Hearing tests
Facts	C1 Hearing tests <*hasNarrowerTerm*> Audiometry C2 Hearing tests <*hasNarrowerTerm*> Acoustic impedance tests
Rules	Document X <*dealsInclusivelyWith*> Subject Y IF Document X <*dealsWith*> Subject Y Document X <*dealsInclusivelyWith*> Subject Y IF Subject Y <*hasNarrowerTerm*> Subject Z AND Document X <*dealsWith*> Subject Z
Answer	Document 1 (due to fact B1 document 1 <*dealsWith*> Acoustic impedance tests) Document 2 (due to fact B7 document 2 <*dealsWith*> Hearing tests) Document 4 (due to fact B11 document 4 <*dealsWith*> Audiometry)

The human reader can assimilate hierarchy facts better in a linear arrangements as shown below:

Hierarchy excerpt from Medical Subject Headings

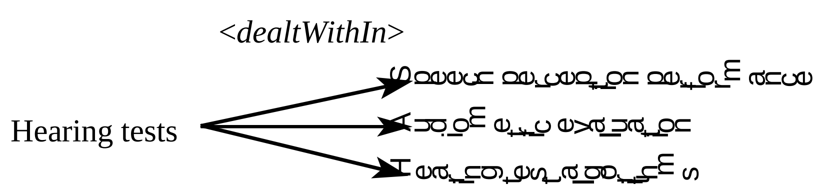
E1	**Diagnosis**
E1.276	. Diagnosis, otorhinolaryngologic
E1.276.299	. . Diagnosis, ear
E1.276.299.375	. . . **Hearing tests**
E1.276.299.375.100	. . . . **Acoustic impedance tests**
E1.276.299.375.297	. . . . **Audiometry**
E1.276.299.375.297.45	. . . . . Audiometry, evoked response
E1.276.299.375.297.92	. . . . . Audiometry, pure-tone
E1.276.299.375.297.105	. . . . . Audiometry, speech
E1.276.299.375.297.105.89	. . . . . . Speech discrimination tests
E1.276.299.375.297.105.902	. . . . . . Speech reception threshold test
E1.276.299.375.330	. . . . Dichotic listening tests
E1.276.299.375.570	. . . . Recruitment detection (audiology)
E1.276.299.816	. . . Vestibular function tests
E1.276.299.816.250	. . . . Caloric tests
E1.276.299.816.435	. . . . Electronystagmography
E1.276.591	. . Laryngoscopy
E1.276.660	. . Nasal provocation tests

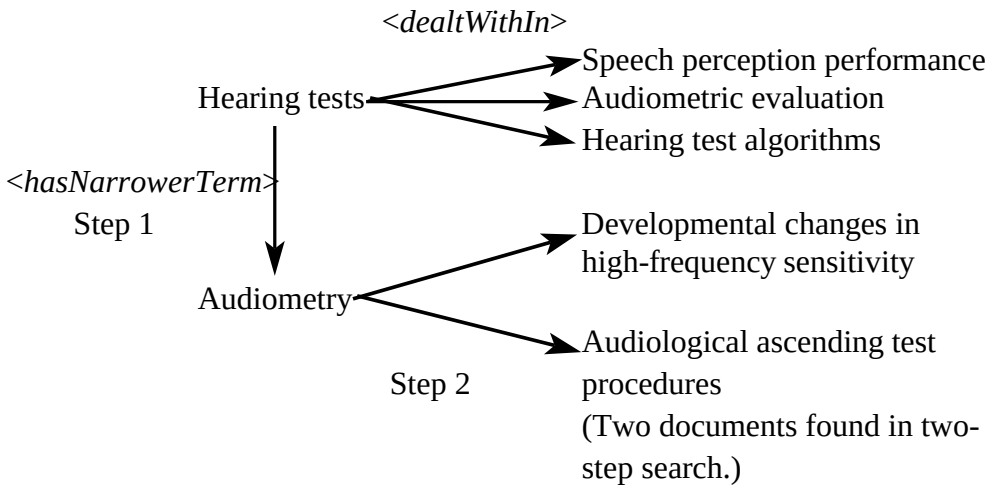
Note: The term numbers (also called codes or notations) make the connection between an alphabetical index and the hierarchy listing.

Types of information systems from simple to complex (and more useful)

simple: System just finds data, user does all the processing needed to find an answer, often adding knowledge from other sources

to complex: System does a lot of processing for the user, provides final answers / problem solutions

Information systems by extent of processing	Information systems differ in the **extensiveness of their knowledge base** (or database) and the **intensity of information processing** to find or create an answer. The more extensive the knowledge base and the more intensive information processing, the more useful are the answers the system can give and the easier is interaction with the system. **Plain retrieval systems** vs. **knowledge-based systems**, **intelligent information systems**, **expert systems** The term **decision support system** is also used, particularly in connection with systems that use simulation and modeling to support business decisions.
Types of information processing	• inferential reasoning • mathematical computations • statistical analysis • simulation and modeling • neural networks • genetic algorithms
Plain information retrieval or database system	A plain information retrieval or database system finds answers from statements that exist ready-made in the database. Another way of saying this: A plain IR system uses one-step linkages. Example: bibliographic IR system Question: Find documents dealing with Hearing tests Query: Document X <*dealsWith*> Hearing tests Answers (document title) Speech perception performance <*dealsWith*> Hearing tests Audiometric evaluation <*dealsWith*> Hearing tests Hearing test algorithms <*dealsWith*> Hearing tests  Could also use links from words in the text or from person who is author.

Expert system	An **expert system** uses a chain of inferences relying on many types of data concerning many types of objects/entities, for example: ! Prescription of drugs is based on data about the illness to be treated, the effectiveness of drugs against certain illnesses, contra-indications of drugs, and other conditions of the patient. ! Expert system for college choice. Such a system starts by simply comparing the criteria entered by the user with the corresponding data about the colleges – simple retrieval. But such a system would also consider user characteristics (such as grades and test scores) and compare them with the admissions standards of the college – qualified by subject applied for and other relevant factors – and thus arrive at a probability of admission. Or it would use data about alumni who are relatives of the user - if these data are available. ! Inclusive (explode) searching in MEDLINE uses data on the hierarchical relationships between descriptors in addition to the data about document-descriptor linkages. So it does combine two types of data to arrive at retrieval results and could therefore be called an expert system. But inclusive searching is a borderline case, and MEDLINE is not commonly seen as an expert system (even though it mimics an expert librarian).  ! There is no sharp boundary between ordinary information systems and expert systems (also called knowledge-based systems). The more different types of facts are in the system and the more inference (combination of different types of facts) is used in deriving answers, the more expert the system is. Medline would not normally be considered an expert system, but it is capable of inclusive searching, thus it uses knowledge about concept relationship just as a knowledgeable reference librarian would.

Characteristics of a good information system

- Adapts to the special needs of the user and the specific situation.
- Interprets requests (including understanding natural language) and asks user for clarification when needed. Engages users in a dialog to clarify requests.
- Processes raw data and gives answers that are directed toward the solution of the user's problem or a solution itself, saving the user the considerable effort required for assimilating and processing raw data.
- Asks for more information if it is needed to derive a good answer.
- Gives answers in easily-understood format.
- Gives reasons for suggested problem solutions, explains its reasoning.
- Assists in knowledge acquisition, for example by extracting facts from text.
- Learns.

Advanced ideas to ponder

Interrelatedness of knowledge
- Inference relationships
- Contradictory knowledge

More input/output

Understanding graphical representation, receiving instrument-generated data
Generating language and graphics.

Expert system examples (under construction)

Expert systems can give us ways to build solutions to real problems. Examples of things that an expert system might do:

· Diagnosis and advice (medical diagnosis and advice, automotive diagnosis and advice, skin care and cosmetics, color combinations, ...).
http://easydiagnosis.com/
OSHA eTools and: www.osha.gov/dts/osta/oshasoft/index.html

· Troubleshooting techniques for machinery (cars, phones, household appliances etc), a variation on 1.

· Identifying plants, fish, insects etc.

· Selecting foods for particular occasions.

· Support for making a decision or choice, for example choosing a music CD based on ones you enjoy or hate.

· Working out the best way to do some task (for example, what is the best way to get from Kings Meadows to Invermay on a Friday night?)

· Making a decision on a mortgage product (consumer) or on approving a mortgage (bank)
www.bankrate.com/brm/mortgage-advisers/home.asp

· Making a decision on what school to apply to (student) or what students to admit (university/college)
http://ieeexplore.ieee.org/iel5/8934/28293/01265222.pdf?arnumber=1265222
Related: Choice of major http://findarticles.com/p/articles/mi_m0FCR/is_4_36/ai_96619963

· Configuring a computer or other machinery

· Applying cataloging rules

(From www.education.tas.gov.au/itproject/topics/expertsystems/expertsystems.htm)

For more information: **www.aaai.org/AITopics/html/expert.html**
www.generation5.org/content/2005/Expert_System.asp

Supplement for Lecture 2.1

Advanced objective

3 Solidify the understanding of the approaches to knowledge representation as a basis for evaluating knowledge representation schemes

2.2 Objectivist vs. organism-centered view of categories.

A balanced view. Elaboration

Scientists believe that visual information is processed along two pathways in the brain, one which specializes in spatial information and coordinating vision with action, and a second which identifies objects (Anderson, 2005, p. 41). The first pathway ties directly into the emotional center of the brain which controls the flight or fight response. This pathway is much faster than the second, such that all visual information is filtered through an affective response before being evaluated cognitively (Barry, 2005, p. 45-62). The brain attempts to match what has been seen with previous templates, patterns, or features in order to identify what was viewed and assign an appropriate response (Anderson, 2005, p. 48-58)/|Because of this, human visual perception does. not duplicate reality.

> ... what our eye register is not a picture of reality as it is. Rather our brains combine information from our eyes with data from our other sense, synthesize it, and draw on our past experience to give us a workable image of our world. This image orients us, allows us to comprehend our situation, and helps us to recognize significant factors within it....The visual world, then, is an interpretation of reality but not reality itself. It is an image created in the brain, formed by an integration of immediate multi-sensory information, prior experience, and cultural learning (Barry, 1997, p. 15).

Because the human brain processes visual information based on a variety of unique factors, meaning often varies among individuals and even over time for the same individual.

Barry, M. (1997). Image Intelligence: Perception, Image, and Manipulation in Image Communication. New York: State University of New York Press.

From Rachael Bradley dissertation, p. 26

2.4 **Basic level categories** (Eleanor Rosch)

Optional

Further Quotes here are from Rosch, Eleanor. *Classification of real-world objects: Origins and representation of cognition.* Johnson-Laird and Wason, eds. *Thinking*. 1977)

Importance - some applications in information systems:

- Classification for children's collections
- Easiest level of specificity in indexing
- Book at right level of specificity for reader
- Medical concepts known to health consumers (?)

"In so far as categorization occurs to reduce the infinite differences between stimuli to behaviorally and cognitively usable proportions, two opposing principles of categorization are operative:

(a) On the one hand, it is to the organism's advantage to have each classification as rich in information as possible. This means having as many properties as possible predictable from knowing any one property (which, for humans, includes the category name), a principle which would lead to formation of large numbers of categories with the finest possible discriminations between categories.

(b) On the other hand, for the sake of reducing cognitive load, it is to the organism's advantage to have as few classifications as possible, a principle which would lead to the smallest number of and most abstract categories possible.

We believe that the basic level of classification, the primary level at which 'cuts' are made in the environment, is a compromise between these two levels; it is the most general and inclusive level at which categories are still able to delineate real-world correlational structures." (p. 213)

How basic level categories apply in information architecture (slides):

http://www.kapsgroup.com/presentations/Semantic%20Technology%20-%20Basic%20Categories.ppt

Supplement for Lecture 2.2

Advanced objective

3 Solidify the understanding of the approaches to knowledge representation as a basis for evaluating knowledge representation schemes

4 Some criteria for describing and evaluating knowledge representations (advanced)

These criteria can be applied

- to the syntax (the format of knowledge representation);
- to the conceptual data schema (entity types and relationship types);
- to the vocabulary (entity values).

Distinguish between domain-independent vocabulary and domain-dependent vocabulary. For example, in medicine such terms as **asthma** and **prednisone** are domain-dependent (domain-specific) while such terms as **cost-benefit analysis** and **triage** are domain-independent (general)

Completeness, expressiveness, detail (subdivided by type of knowledge)

- Extensibility - can easily add new types of knowledge

Parsimony of syntax and of vocabulary - use small number of syntactic constructs and of entity and relationship types

Modularity

- In a modular system, small pieces of knowledge can be added to the knowledge base without changing what is already there

Compactness / redundancy

- In a compact system, knowledge that can be inferred or derived is not stored but produced on the fly as needed, which may take time. In a redundant system, inferable knowledge is stored explicitly; this may save time but does take up space. An additional problem is that when knowledge changes stored inferred knowledge may no longer be true; the system has to watch out for that (truth maintenance).

Ease of processing by people or by computer programs

- Ease of producing a knowledge base
- Ease of writing knowledge items
- Support for knowledge elicitation, support for association
  - Consistency checks
  - Plausibility checks
- Ease of retrieval
- Ease of reading
- Ease of reasoning, drawing inferences by deduction and induction

The material at the end of Lecture 1.2 is related.

Supplement for Lecture 5.1

4.2 Further elaboration of data structures (Advanced, ➜LIS 506 Information Technology)

Relational databases: Storage in tables (example: University Database in Chapter 3)

Each table contains all the statements that use the same relationship type; statements pertaining to one entity value are distributed over many tables. In retrieval, data can be combined in many ways. The system gives equal consideration to the user who wants to know everything about a document, including the person who authored it, and to the user who wants to know everything about a person, including the documents he authored.

"Flat file" databases: Storage in records

As discussed in Lecture 4.2 (Organizing Information, Section 9.2), a record assembles the information about one entity value - the various statements that pertain to that entity value. Records are needed for eliciting input and for presenting output. Often, storage is also based on records. With storage records, statements pertaining to one entity value are all in one place, while statements using the same relationship type are distributed over many records. Storage by records introduces a perspective or focus: If data are assembled in document records, the data structure gives more consideration to the user who wants to know everything about a document; the linkage between a document and the person who authored it is stored in the document record. If, on the other hand, data are assembled in person records, the data structure gives more consideration to the user wanting to know everything about a person; the linkage between a document and the person who authored it is stored in the person record. By storing the same information twice, both users can be accommodated.

See also the example on bibliographic data in MARC records (flat file) and as a collection of

Object-oriented databases are based on frames with hierarchical inheritance (see Lecture 2.2). They are closer to the record model than to the relational model.

Searching printed indexes vs. searching by computer.

Division of labor between system and user: Degree of order and amount of information presented in search output (See example 13 from *Design of an integrated information structure interface. Prologue*.)

Supplement for Lecture 5.2a

Elaboration of text types adapted from Beaugrande *Text, discourse, and process,* VII.1.8

Descriptive	The text revolves around **object and situation concepts**, about which statements are made through links in multiple directions. The link types of *state, attribute, instance, and specification* are frequent. The surface text reflects a corresponding density of *modifier* dependencies. The most commonly applied global knowledge pattern is *the frame*.
Argumentative	The text revolves around **entire propositions** which are assigned values of truthfulness and give reasons for considering beliefs as facts; often there is an opposition between propositions with conflicting value and truth assignment. The link types of *value, significance, cognition, volition, and reason* are frequent. The surface text contains a density of evaluative expressions. The most commonly applied global knowledge pattern is the plan whose goal state is the inducement of shared beliefs.
Didactic	The text revolves around a **topic or theme about which the receiver is to learn something**, that is, integrate new objects and relationships into her memory. The text must present the subject via a process of gradual integration, because the receiver does not yet have the matchable knowledge spaces that a scientific text would require. Therefore, the linkages of established facts are problematized (put into question) and then de-problematized.
Narrative	The text revolves around the main **event and action concepts** which are arranged in an *ordered directionality* of linkage. The link types of *cause, reason, enablement, purpose,* and *time proximity* are frequent). The surface text reflects a corresponding density of *subordinative* dependencies. The most commonly applied global knowledge pattern is the *schema*. (Freedle and Hale (1979) show that a narrative schema, once learned, can easily be transferred to the processing of a descriptive text on the same topic.)
Conversational	The text has an especially **episodic and diverse range of sources for admissible knowledge**. Less emphasis on expanding current knowledge of the participants than for the other text types. The surface organization assumes a characteristic mode because of the changes of speaking turn.

Literary	The text revolves around **alternatives to matchable patterns of knowledge about the accepted real world**. The intention is to motivate, via contrasts and rearrangements, some new insights into the organization of the real world. From the standpoint of processing, the linkages within real-world events and situations is PROBLEMATIZED, that is, made subject to potential failure, because the text-world events and situations may (though they need not) be organized with different linkages. (*Problematize* = put into question, consider as uncertain, therefore problematic.) The effect is an increased *motivation* for linkage on the side of the text producer and increased focus for linkage on the side of the receiver. This problematized focus sets even "realistic" literature (reaching extremes in "documentary" art) apart from a simple report of the situations or events involved: the producer intends to portray events and situations as *exemplary* elements in a framework of *possible alternatives.* In **poetic texts**, the alternativity principle is extended to the *interlevel mapping of options*, e.g. sounds, syntax, concepts/ relations, plans, and so on. In this fashion, both the organization of the real world and the organization of discourse about that world are problematized, and the resulting insights can be correspondingly richer.
Scientific	The text revolves around an **optimal match with the accepted real world** unless there are explicit signals to the contrary (e.g., a disproven theory). Rather than alternative organization of the world (as in literary text, see above), a more exact and detailed insight into the established organization of the real world is intended. In effect, the linkages of events and situations are eventually *de-problematized* via statements of causal necessity and order.

Lecture 5.2b Text analysis overview and examples. Supplement

In-class exercises and examples illustrating the importance of text analysis through several linguistic techniques

2 Extracting data through slot-filling in frames: examples

Extracting data from pesticide reports

Pesticide frame

Slot	**Instructions: What to look for to find slot fillers**
Substance	a term that designates a substance
Pest fought	the name of an organism that can be a pest or the name of a disease
Crop or livestock	the name of a useful plant or animal
When applied	the name of a season or a term indicating weather condition
Dosage	a symbol for mass, such as *pound*, *g*, *kg* and the number preceding. Also look for *per* or *for each*
Route of administration	a term such as *spray*, *work into the soil*

3 Exacting data from text, especially importance of **resolving anaphoric references**

Contact Dermatitis-Irritant and Allergic

Contact dermatitis may result from irritants or substances to which an individual has become allergic. Depending upon the source of irritation, the duration or frequency of exposure, and other variables, different uncomfortable changes in the skin occur.

Irritant contact dermatitis occurs when the skin is exposed to a mild irritant-such as detergents or solvents-repeatedly over a long period of time or to a strong irritant, such as acid or alkali, which can cause immediate damage to the skin.

This disorder is an "occupational hazard" for housewives, chemical workers, doctors and dentists, restaurant workers, and others whose work brings them into regular or prolonged contact with **soaps, detergents, chemicals, and abrasives**. *These substances* either erode the protective oily barrier of the skin or injure its surface.

Allergic dermatitis occurs when skin which has been sensitized to a specific substance comes in contact with that substance again. With the exception of poison ivy and poison oak, to which about 70 percent of people become sensitized after first contact, most contact allergies produce sensitivity in only a few people. The most common of these allergies are nickel and other metals, rubber and elasticized garments, dyes, cosmetics (especially nail polish), and leather. But anyone can become sensitized to almost anything, so the search for the offending substance is often tedious and success is sometimes elusive.

In **irritant dermatitis** the **skin** becomes stiff, dry, and tight-feeling. *It* may crack, blister, or become ulcerated. Some itching may accompany mild inflammation, but the fissures and ulcers will be painful, not itchy. Mild irritants cause a progression from reddening and blistering to drying and cracking, while strong irritants cause blistering on contact and then erosion and ulcers.

Allergic dermatitis appears as reddening, followed by blistering and oozing. *In severe cases* there may be swelling of the face, eyes, and genital area. The rash will appear wherever the allergen has touched the skin, either directly or by transference from the hands. However, the palms, soles, and scalp seldom show any reaction. Fluid from the blisters will not spread the disease to other parts of the body or to other people.

There are no tests to determine the cause of **irritant dermatitis**. Finding the source may require persistent and creative detective work on the part of both doctor and patient. Patch tests can often determine or point the way to the allergens responsible for the reaction in **allergic dermatitis**. It may, however, take some sleuthing to find the specific product or products which contain the offending substance.

Preventive measures for irritant dermatitis are easy to define and difficult to carry out. The disease is usually the direct result of the working environment, and adequate protective measures are often impractical, if not impossible, to achieve. To the extent possible, then, it is recommended that the patient take the following precautions:

1. Wear cotton gloves under rubber gloves for all wet work. If gloves are impractical, use a barrier cream to protect the skin. Reapply the cream 2 or 3 times per day and after each handwashing.

Finally, consider this text:

Leukemia

Acute Lymphocytic Leukemia (ALL) and **Chronic Myelogenous Leukemia** (CML) occur in different populations with different symptoms. *The former* primarily affects children under age 5, who often show signs of anemia, fatigue, fever, and bleeding, indicating a depressed functioning of the bone marrow. *The latter* occurs primarily in men between 20 and 50, with symptoms varying from none at all to anemia and general malaise to weight loss, night sweats, fatigue, and an enlarged spleen that may cause discomfort on the left side of the abdomen. *The disease* can develop gradually, almost insidiously. The number of granulocytes is markedly increased, . . .

Application to searching (advanced exploration)

Try searching for some of the noun phrases from example 2 in Google. Just type them in without using quotes. In all cases, a large proportion of the top 100 documents (Web sites, but in Google Scholar also articles) found have the noun phrase in them. So Google must have some mechanism for searching phrases; it may be as simple as giving a document a higher score if the search words are close together.

Sequence also seems to matter. *library school* gets results about evenly divided between library schools and school library (school at all levels, not just K-12, the meaning of school in the phrase *school libraries*)

Try *peer pressure*, *pressure by peers*, and *pressured by peers*
The first two find very similar Web sites, the last finds additional relevant sites

Try *social pressure*
Look-ahead note: While all of the *peer pressure* Web sites are relevant, only a few are found

A system could use noun phrases to disambiguate homonymous and polysemous words, so it would know whether *pressure* means *physical pressure* (as in *vapor pressure*, *water pressure*, *barometric pressure*) and when it means "*mental pressure"* (as in *peer pressure*, *parental pressure*, *social pressure*). Then the user could search for these general concepts, whereas in Google a search for *pressure* returns everything.

5.2c Supplement

Natural language processing (NLP) achieves the purposes listed in *Practical significance* through several techniques

Identifying noun phrases	(in all their variant forms) in document texts and in query statement texts as good indexing terms and search terms, respectively. (Some search engines look for noun phrases in the string of words entered into the query box and rank documents with the noun phrase higher than documents that just have the individual words.) Note the difficulty posed by situations like *information retrieval*, *retrieval of information*, *retrieval of legal information*; looking simply for the string *information retrieval* will give incomplete results. But that is what the above-mentioned search engines most likely do, because the alternatives are (1) still costly syntactic processing of all Web page texts or (2) using proximity operators, which is less precise.
Complete or partial sentence parsing	Note: Emphasis is not so much on the role of a parser identifying a string of words as a well-formed sentence. What really matters is: • identifying the role of each word or group of words in the sentence, which is the basis for determining part of speech of a word (is man used as a noun or a verb?), • identifying noun phrases, • semantic parsing For purposes of simply "understanding" the text, it is even useful if the system can deal with sentences that are not well-formed; in this context, checking for grammaticality is important only insofar as it supports understanding, especially through disambiguation.
Semantic parsing	Disambiguating homonyms, word sense disambiguation (WSD)
Statistical NLP methods	Increasingly used for several functions, replacing or working in combination with formal syntax methods • part-of-speech tagging • summarization • automatic translation • automatic speech recognition

Statistical and formal methods	As we discussed, both statistical analysis and syntactic analysis are used for NLP. Systems differ in the degree to which they rely on these two approaches. All of the purposes listed below are amenable to either approach; automatic summarization of single documents is usually done statistically, multi-document summarization systems and information extraction systems often use at least some syntactic and semantic processing.
Multiple languages	The methods discussed can be applied to any language; of course, each language needs its own syntax and semantics knowledge base. Statistical systems may process a multilingual collection; syntactic-semantic systems usually deal with one language at a time. One could put together many such systems into one package, with a program that can recognize the language of a document sending incoming documents to the appropriate language-specific program.

Examples of statistics-based and NLP-based summarizers
Overview: http://itt.nissat.tripod.com/itt0202/ruoi0202.htm **www.copernic.com/en/products/summarizer/** The MS Word AutoSummarize function on the Tools menu http://domino.research.ibm.com/cambridge/research.nsf/0/74c0a77cbfad5ae585256bf80054b036?OpenDocument

`Example NLP tools, including parsers
This site has many links to NLP tools, nicely classified http://**www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html**

This lecture uses **transition network diagrams** as an example to illustrate parsing. These diagrams are intended as the blueprint for a computer program that could process a document one sentence at a time. Inter-sentence relationships, such as anaphoric reference, would have to be detected in a second phase. We will start with the analysis of noun phrases and then move to simple sentences. A full parsing system would be orders of magnitude more complex.

In-class exercise in parsing: Identification of noun phrases for indexing

P. 145 - 154 The parsing game (take these pages out of your binder)

P. 156 - 171 More detail about the syntactical analysis (look at together with the parsing game)

571 Soergel

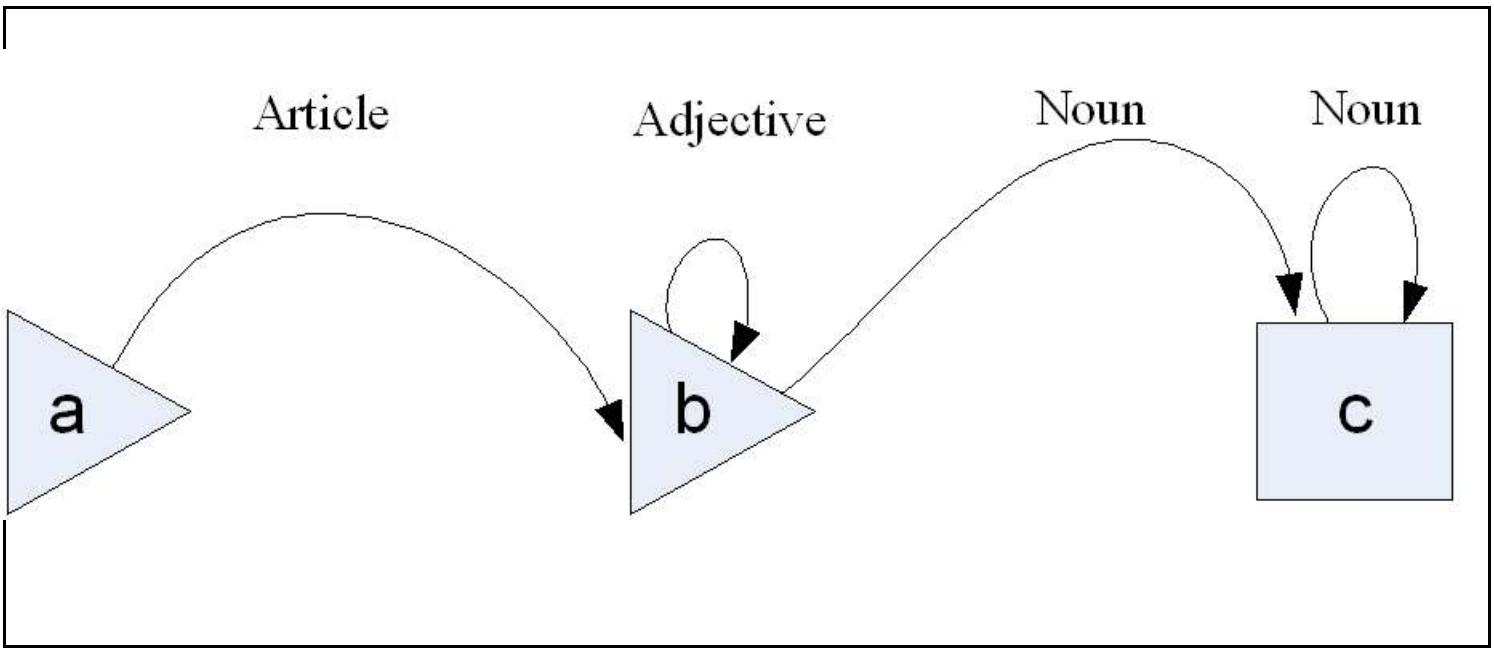
The parsing game

To start, put game piece on a triangle.

Move game piece along the arc corresponding to the next work in the string of words, cross off the word
If you cannot move and there are still words left, you loose.

If you arrive at a square and no words are left, you win.

Simple transition network diagram: noun phrase



Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ dishwasher $_2$

2 $_0$ the $_1$ jolly $_2$ dishwasher $_3$

3 $_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$

4 $_0$ bones $_1$

5 $_0$ regular $_1$ daily $_2$ consumption $_3$

6 *$_0$ daily $_1$ consumption $_2$ regular $_3$

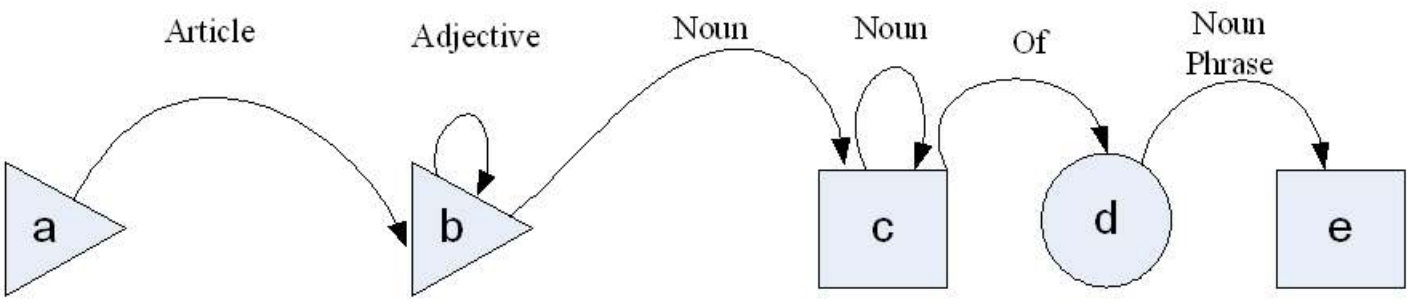
7 $_0$ bone $_1$ mass $_2$

8 $_0$ the $_1$ calcium $_2$ supply $_3$

9 *$_0$ supply $_1$ calcium $_2$

10 $_0$ a $_1$ deficient $_2$ calcium $_3$ supply $_4$

Noun phrase network (1)



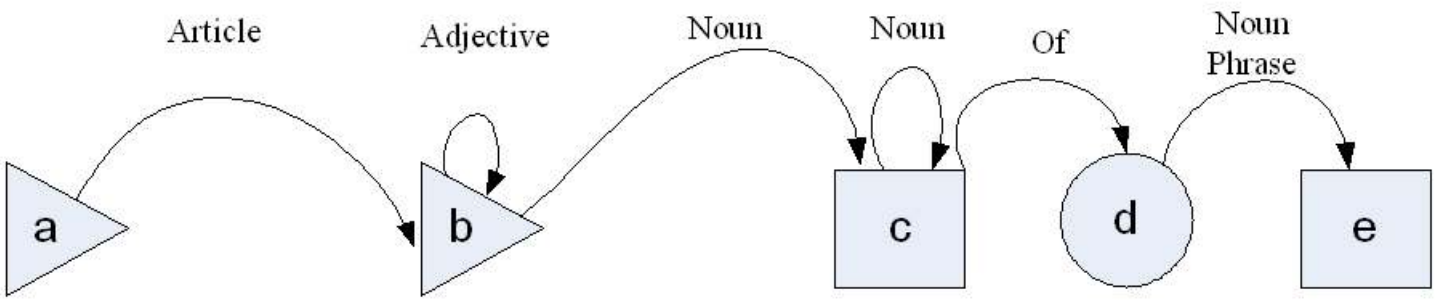
Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$

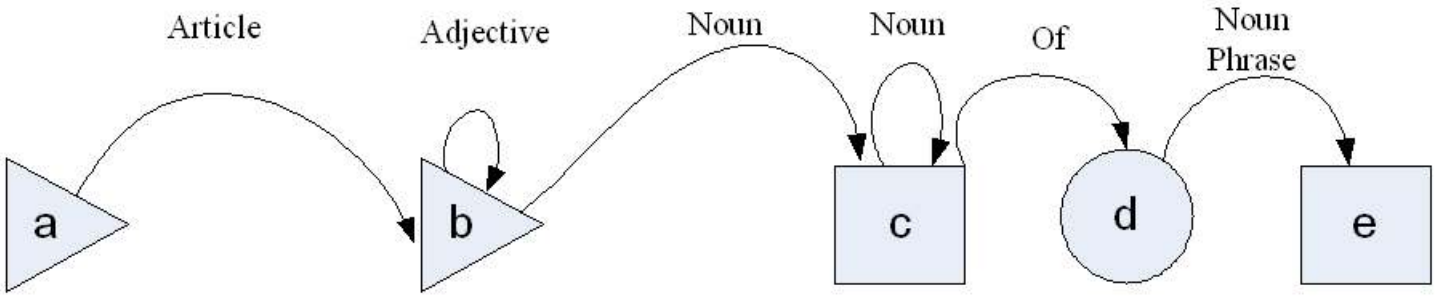
2 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$

3 $_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$ of $_9$ healthy $_{10}$ parents $_{11}$

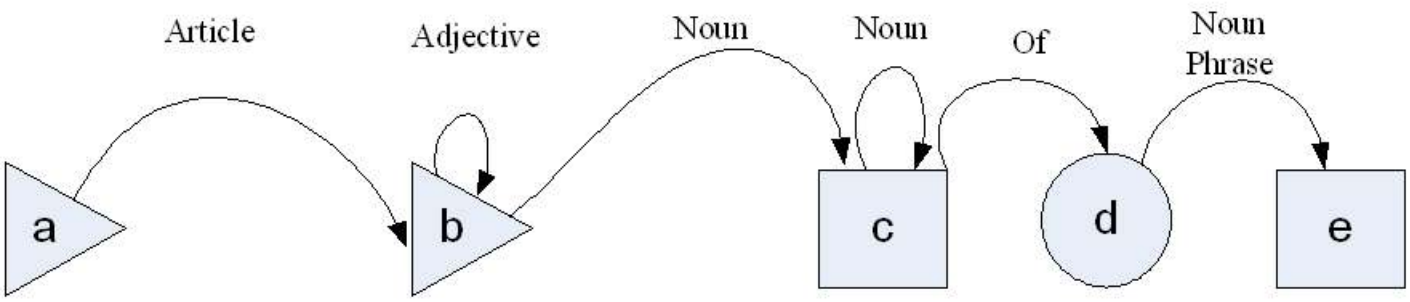
Noun phrase network (2)



Noun phrase network (3)



Noun phrase network (1)



OSTEOPOROSIS

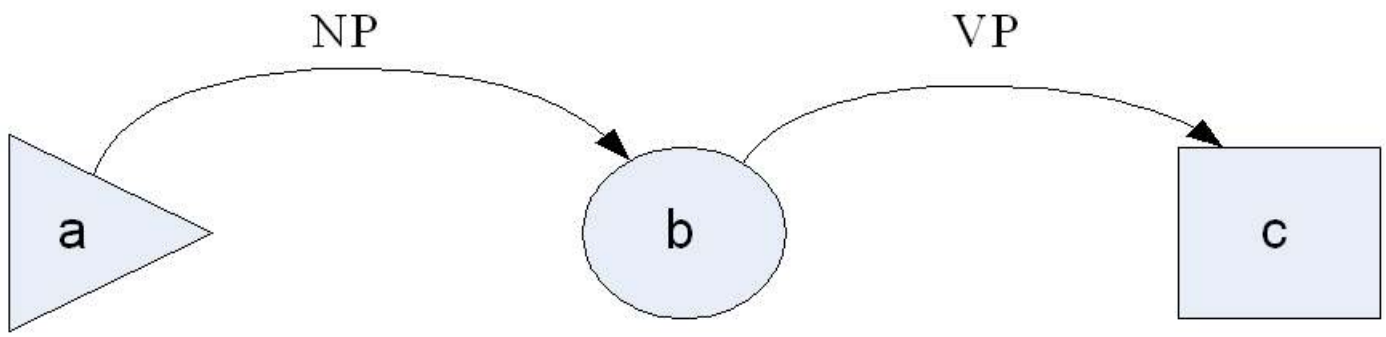
BONES NEED CALCIUM to maintain their strength, hardness, and to stay healthy. Milk, the main source of calcium in the diet, is important for the growing skeletons of children and adolescents as well as the bone-forming cells of adults. Regular daily consumption of at least 1 cup of skim or low-fat milk is essential for adults who want to keep their bones strong and to help prevent osteoporosis, a disease in which the body's bone mass decreases and bones become thin and brittle. Bones weakened by osteoporosis, a disease common to postmenopausal women, are prone to fracture if a person falls.

When calcium enters the body, it is absorbed into the bloodstream. If there is any excess, it is deposited in the end of the bone shafts where it is stored until the body needs to tap this reserve. (Some is also excreted via the kidneys.) When the calcium supply is deficient, the blood must take it back from the bones. If calcium intake remains inadequate over a long period of time, the bones eventually become porous and weak.

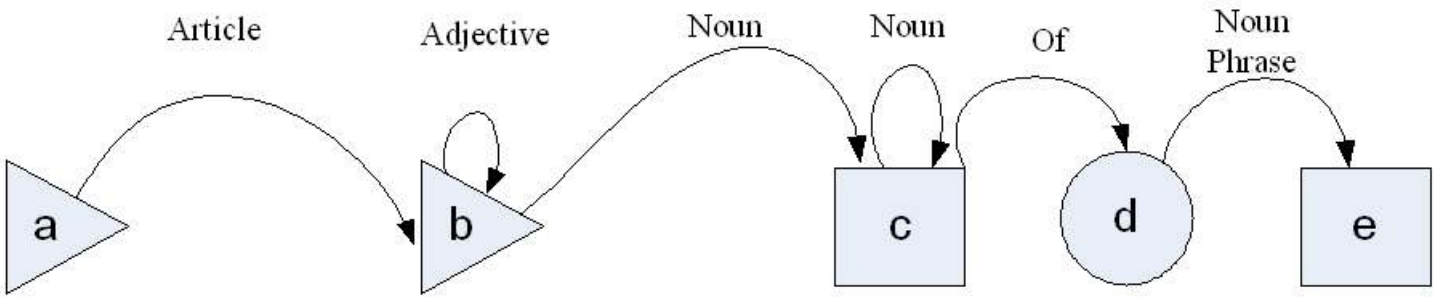
It is not known why calcium loss occurs. That postmenopausal women tend to get osteoporosis points in the direction of a hormonal disorder as estrogen in women of this age falls off sharply. Estrogen therapy is one treatment but its ability to decrease calcium loss may last only several years. Increased calcium intake and exercise are other therapies. The links between lack of exercise and osteoporosis are becoming firmer as research into the causes of this disease progresses.

The disease most frequently affects the spinal column, causing backaches and rounded shoulders. in severe cases, the bone becomes as porous as a sponge and can collapse as a result. Collapsing **vertebrae**, which can cause sudden and sharp backaches, is one reason why elderly people tend to get shorter.

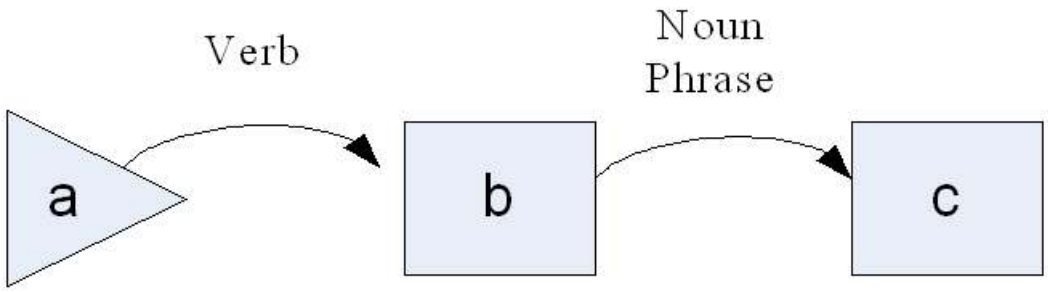
Sentence network



Noun phrase network (1)



Verb phrase network

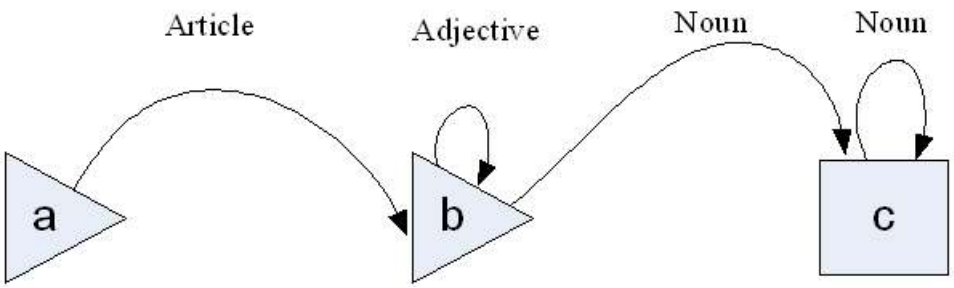


1 $_{0}$ **The** $_{1}$ **green** $_{2}$ **vegetables** $_{3}$ **supply** $_{4}$ **calcium** $_{5}$**.**
2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]
3 The green vegetables supply digestible calcium.
4 The green vegetables supply determines sufficiency of calcium.

Go to next page

Simple transition network diagram: noun phrase

NP



▷ possible start state G possible end state

(Note that there are two possible starting points; if the first does not work, try the second)

Dictionary

a ART	dishwasher N
bone N	jolly ADJ
bones N	mass N
calcium N	regular ADJ
consumption N	supply N
daily ADJ	the ART
deficient ADJ	white ADJ

Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ dishwasher $_2$

2 $_0$ the $_1$ jolly $_2$ dishwasher $_3$

3 $_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$

4 $_0$ bones $_1$

5 $_0$ regular $_1$ daily $_2$ consumption $_3$

6 *$_0$ daily $_1$ consumption $_2$ regular $_3$

7 $_0$ bone $_1$ mass $_2$

8 $_0$ the $_1$ calcium $_2$ supply $_3$

9 *$_0$ supply $_1$ calcium $_2$

10 $_0$ a $_1$ deficient $_2$ calcium $_3$ supply

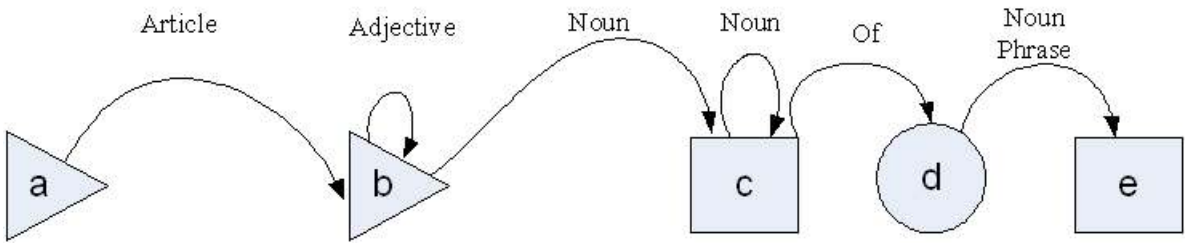
Step-by-step trace of the parsing process

From pos	From state	Arc tried	segment (word) processed	To state	To pos	Comment
$_0$ the $_1$ dishwasher $_2$						
0	a	ART	the	b	1	
1	b	NOUN	dishwasher	**c**	**2**	end state, all words used = **success**
$_0$ the $_1$ jolly $_2$ dishwasher $_3$						
0	a	ART	the	b	1	
1	b	ADJ	jolly	b	2	
2	b	NOUN	dishwasher	**c**	**3**	**success**
$_0$ the $_1$ jolly $_2$ white $_3$ dishwasher $_4$						
0	a	ART	the	b	1	
1	b	ADJ	jolly	b	2	
2	b	ADJ	white	b	3	
3	b	NOUN	dishwasher	**c**	**4**	**success**
$_0$ regular $_1$ daily $_2$ consumption $_3$						
0	a	ART	~~regular~~	a	0	Try next possible start state, namely b.
0	b	ADJ	regular	b	1	
1	b	ADJ	daily	b	2	
2	b	NOUN	consumption	**c**	**3**	**success**
***$_0$ daily $_1$ consumption $_2$ regular $_3$**						Not a good noun phrase
0	a	ART	~~daily~~	a	0	No arc to follow
0	b	ADJ	daily	b	1	

1	b	NOUN	consumption	c	2	
2	c		regular	**c**	2	No arc to follow, **failure**

Complex transition network diagram: noun phrase

NP



▷ possible start state ◯ intermediate state G possible end state

(Note that there are two possible starting points; if the first does not work, try the second)

Dictionary

a ART	main ADJ
bone N	mass N
bones N	of PREP
calcium N	parents N
children N	parts N
consumption N	regular ADJ
daily ADJ	skeleton N
deficient ADJ	small ADJ
dishwasher N	source N
growing ADJ	supply N
healthy ADJ	the ART
jolly ADJ	white ADJ

Sample noun phrases (by general linguistic convention, * means syntactically incorrect)

1 $_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$

2 $_{0}$ the $_{1}$ growing $_{2}$ skeleton $_{3}$ parts $_{4}$ of $_{5}$ healthy $_{6}$ small $_{7}$ children $_{8}$

3 $_{0}$ the $_{1}$ growing $_{2}$ skeleton $_{3}$ parts $_{4}$ of $_{5}$ healthy $_{6}$ small $_{7}$ children $_{8}$ of $_{9}$ healthy $_{10}$ parents $_{11}$

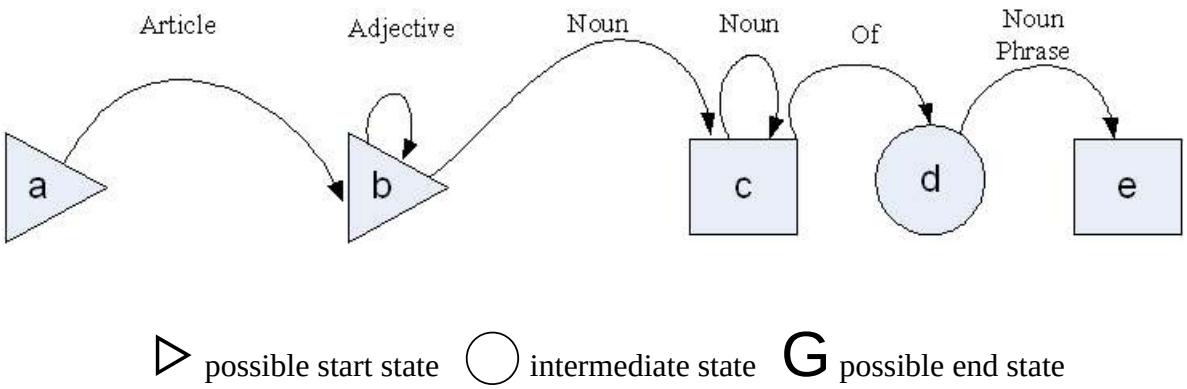
From pos	From state	Arc tried	segment (word) processed	To state	To pos	comment
$_0$ the $_1$ main $_2$ source $_3$ of $_4$ calcium $_5$						
0	a	ART	the	b	1	
1	b	ADJ	main	b	2	
2	b	NOUN	source	c	3	
3	c	OF	of	d	4	
4	d	NP	calcium	**e**	**5**	NP network called again, single noun is a noun phrase **success**

From pos	From state	Arc tried	segment (word) processed	To state	To pos	comment
$_0$ the $_1$ growing $_2$ skeleton $_3$ parts $_4$ of $_5$ healthy $_6$ small $_7$ children $_8$						
0	a	ART	the	b	1	
1	b	ADJ	growing	b	2	
2	b	NOUN	skeleton	c	3	
3	c	NOUN	parts	c	4	
4	c	OF	of	d	5	
5	d	NP	healthy small children	**e**	**8**	NP network called again, this sequence is a noun phrase **success**

Note: These two examples give a first inkling of nesting transition network diagrams. Here we use the NP diagram to process a sequence of words inside a noun phrase that is itself being analyzed with a NP diagram. Here this nesting is treated very informally; examples to follow will demonstrate the exact process.

NP

Complex transition network: noun phrase



Dictionary

a ART
adolescents N
adults N
blood N
bloodstream N
body N
bone N
bone-forming ADJ
bones N
brittle ADJ
calcium N
children N
common ADJ
consumption N
cup N
daily ADJ
deficient ADJ
diet N
disease N
dishwasher N
essential ADJ
excess N
fracture N
growing ADJ
hardness N
healthy ADJ
important ADJ
inadequate ADJ
intake N
jolly ADJ
kidneys N
low-fat ADJ
main ADJ
mass N
milk N
need V N
osteoporosis N
person N
postmenopausal ADJ
prone ADJ
regular ADJ
reserve V N
shafts N
skeletons N
source N
strength N
strong ADJ
supply V N
the ART
thin ADJ
weakened ADJ
white ADJ
women N

Apply the complex transition network and the enlarged dictionary to the identification of noun phrases in the following text.

OSTEOPOROSIS

BONES NEED CALCIUM to maintain their strength, hardness, and to stay healthy. Milk, the main source of calcium in the diet, is important for the growing skeletons of children and adolescents as well as the bone-forming cells of adults. Regular daily consumption of at least 1 cup of skim or low-fat milk is essential for adults who want to keep their bones strong and to help prevent osteoporosis, a disease in which the body's bone mass decreases and bones become thin and brittle. Bones weakened by osteoporosis, a disease common to postmenopausal women, are prone to fracture if a person falls.

When calcium enters the body, it is absorbed into the bloodstream. If there is any excess, it is deposited in the end of the bone shafts where it is stored until the body needs to tap this reserve. (Some is also excreted via the kidneys.) When the calcium supply is deficient, the blood must take it back from the bones. If calcium intake remains inadequate over a long period of time, the bones eventually become porous and weak.

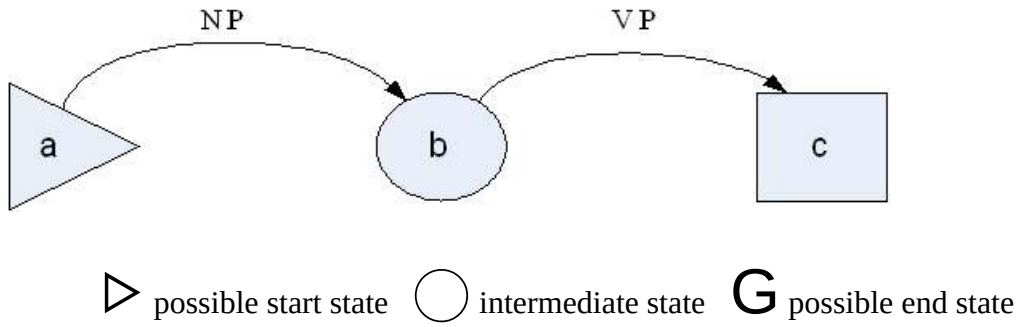
It is not known why calcium loss occurs. That postmenopausal women tend to get osteoporosis points in the direction of a hormonal disorder as estrogen in women of this age falls off sharply. Estrogen therapy is one treatment but its ability to decrease calcium loss may last only several years. Increased calcium intake and exercise are other therapies. The links between lack of exercise and osteoporosis are becoming firmer as research into the causes of this disease progresses.

The disease most frequently affects the spinal column, causing backaches and rounded shoulders. in severe cases, the bone becomes as porous as a sponge and can collapse as a result. Collapsing **vertebrae**, which can cause sudden and sharp backaches, is one reason why elderly people tend to get shorter.

Parsing of sentences: The sentence network outlines a grammar for simple sentences.

Sentence network

S



NP

means: apply the noun phrase parse transition network

	Dictionary
body N calcium N determines V digestible ADJ green ADJ	sufficiency N supply V, N the ART to PREP vegetables N

Sentences

1 $_0$ **The** $_1$ **green** $_2$ **vegetables** $_3$ **supply** $_4$ **calcium** $_5$**.**
2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]
3 The green vegetables supply digestible calcium.
4 The green vegetables supply determines sufficiency of calcium.

Trace of a sentence parse

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium. $_5$

	From pos	From state	Segment processed	To state	To pos
	0	**S^0 a**	**? (consult NP)**	**?**	**?**
			Magic. Result:		
	0	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
	3	**S^0 b**	**? (consult VP)**	**?**	**?**
			Magic. Result:		
	3	**S^0 b**	**supply calcium**	**S^0 c**	**5**
				Success: End state of S, end of word list	

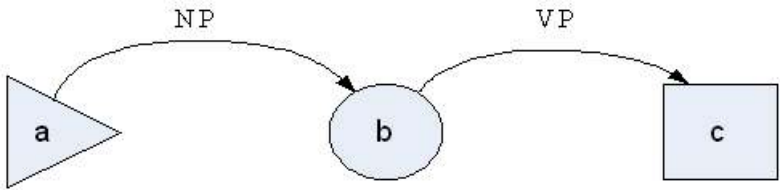
Result: An analysis of the sentence structure, a sentence diagram.

{*S*
[*NP* the green vegetables]
[*VP* supply calcium**]**
}

Parsing of sentences: The three transition networks define a grammar for simple sentences.

Sentence network

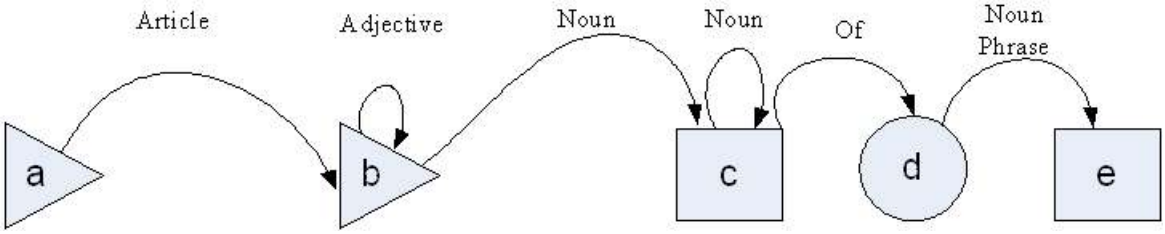
S



▷ possible start state ◯ intermediate state G possible end state

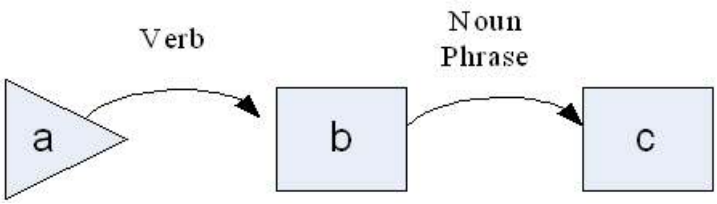
Noun phrase network

NP



Verb phrase network

S



Dictionary	
body N	sufficiency N
calcium N	supply V, N
determines V	the ART
digestible ADJ	to PREP
green ADJ	vegetables N

1 $_0$ **The** $_1$ **green** $_2$ **vegetables** $_3$ **supply** $_4$ **calcium** $_5$**.**

2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]

3 The green vegetables supply digestible calcium.

4 The green vegetables supply determines sufficiency of calcium.

Trace of a sentence parse (Arcs from transition network can be inferred)

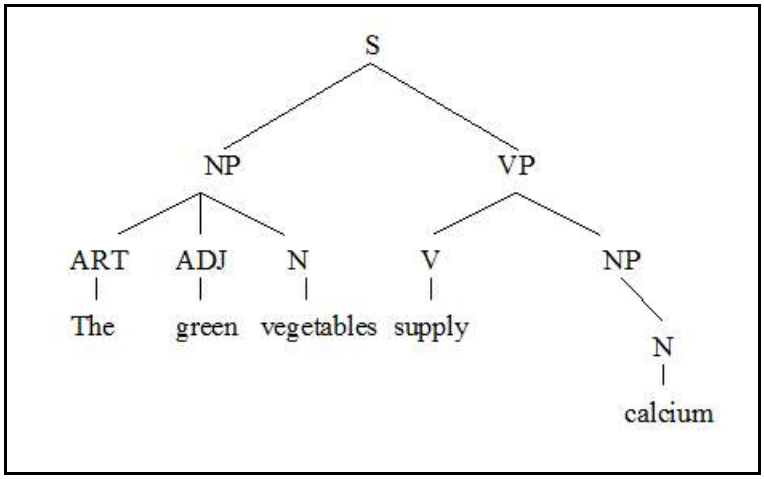
$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ calcium. $_5$

Step	From pos	From state	Segment	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	green	NP^1 b	2
④	2	NP^1 b	vegetables	NP^1 c	3
⑤	**0**	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	supply (V)	VP^1 b	4
⑧	4	VP^1 b	? (consult NP)	?	?
⑨	4	NP^2 a	calcium (*does not work, try starting at b*)	NP^2 a	4
⑩	4	NP^2 b	calcium	NP^2 c	5
1①	4	VP^1 b	calcium	VP^1 c	5
1②	**3**	**S^0 b**	**supply calcium**	**S^0 c**	**5**
1③	**Success:** End state of S, end of word list				

Superscript indicates the nesting depth

Result : **{***S*
[*NP* (*ART* the) (*ADJ* green) (*N* vegetables)**]**
[*VP* (*V* supply) (*NP* (*N* calcium)**]**
}

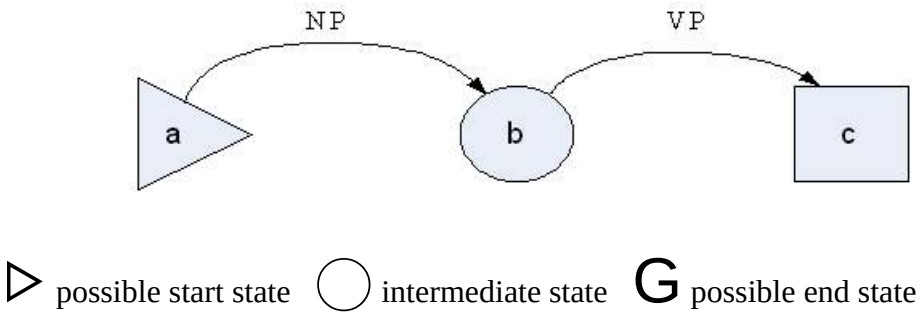
Parse tree representation:



Parsing of sentences: The three transition networks define a grammar for simple sentences.

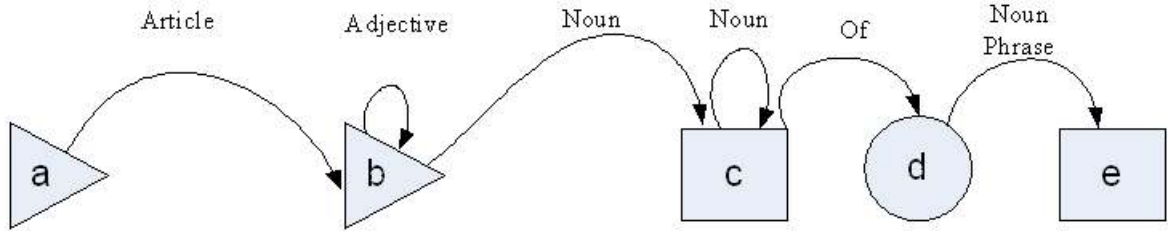
Sentence network

S



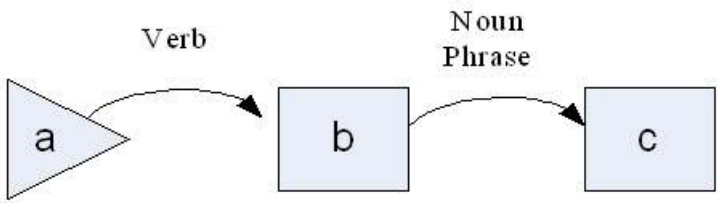
Noun phrase network

NP



Verb phrase network

S



Dictionary	
body N	sufficiency N
calcium N	supply V, N
determines V	the ART
digestible ADJ	to PREP
green ADJ	vegetables N

1 The green vegetables supply calcium

2 The green vegetables supply calcium to the body. [Not recognized by our simplistic parser.]

3 The green vegetables supply digestible calcium.

4 $_{0}$ **The** $_{1}$ **green** $_{2}$ **vegetables** $_{3}$ **supply** $_{4}$ **determines** $_{5}$ **sufficiency** $_{6}$ **of** $_{7}$ **calcium.** $_{8}$

Trace of a sentence parse with backtracking

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ determines $_5$ sufficiency $_6$ of $_7$ calcium. $_8$

Step	From pos	From state	Segment processed	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	green	NP^1 b	2
④	2	NP^1 b	vegetables	NP^1 c	3
⑤	**0**	**S^0 a**	**the green vegetables**	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	supply (V)*	VP^1 b	4
⑧	4	VP^1 b	? (consult NP)	?	?
⑨	4	NP^2 a	determines (*does not work, try starting at b*	NP^2 a	4
⑩	4	NP^2 b	determines (*does not work)*	NP^2 b	4
			Dead end, backtrack to *		
			Dead end, backtrack to *		
1①	3		**Backtrack, continue NP with supply as Noun**		?
	3	NP^1 c	supply (N)	NP^1 c	4
1②	**0**	**S^0 a**	**the green vegetables supply**	**S^0 b**	**4**
1③	**4**	**S^0 b**	**? (consult VP again)**	**?**	**?**
1④	4	VP^1a	determines	VP^1 b	5
1⑤	5	VP^1 b	? (consult NP)	?	?
1⑥	5	NP^2 a	sufficiency (*does not work, try starting at b)*	NP^2 a	5
1⑦	5	NP^2 b	sufficiency	NP^2 c	6
1⑧	6	NP^2 c	of	NP^2 d	7
1⑨	7	NP^2 d	? (consult NP)	?	?
20	7	NP^3 a	calcium (*does not work, try starting at b)*	NP^3 a	7
2①	7	NP^3 b	calcium	NP^3 c	8
2②	7	NP^2 d	calcium	NP^2 e	8
2③	4	VP^1 b	sufficiency of calcium	VP^1 c	8
2④	**4**	**S^0 b**	**determines sufficiency of calcium**	**S^0 c**	**8**
			Success: End state of S, end of word list		

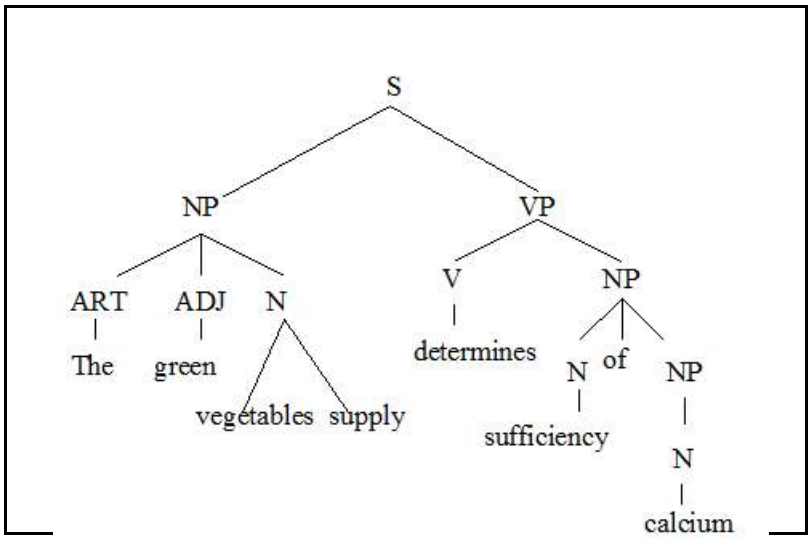
* Backtrack point Superscript indicates the nesting depth

Result: An analysis of the sentence structure, a sentence diagram.

$_0$ The $_1$ green $_2$ vegetables $_3$ supply $_4$ determines $_5$ sufficiency $_6$ of $_7$ calcium. $_8$

{*S* ^{0}
[*NP*^{1} **(***ART*^{1} the**) (***ADJ*^{1} green**) (***N* ^{1} vegetables**) (***N*^{1} supply**)]**
[*VP*^{1} **(***V*^{1} determines**) (***NP*^{2} (*N*^{2} sufficiency) (of^2) (*NP*^{3} (*N*^{3} calcium)))**]**
}

Parse tree representation:



Other example of backtracking:

Compare *The old man cried.* with *The old man the ship.*

Hypothesis: Sentences that do not require backtracking in parsing are easier to read.

Example where backtracking makes reading difficult:

> "Any broadening of the government's role in health risks encouraging employers to give up providing health coverage for employees."
> (Editorial in the Washington Post 1999-7-30)

In a brief search of just the Web I could not find specific research on this. The following lecture materials deal with the issue in general

www.rci.rutgers.edu/~cfs/305_html/Understanding/Understanding_toc.html
(from course Computation and Cognition
www.rci.rutgers.edu/~cfs/472_html/home472.html

The following thesis deals with the problem of if and how people use syntax parsing in understanding sentences. It cites some previous work that found that people take longer in processing syntactically incorrect sentences even if they are not consciously aware of the incorrectness.
http://cognition.iig.uni-freiburg.de/team/members/konieczny/publ/DissLars.pdf

Parser evaluation

Word sequences that will not be recognized as sentences by our very simple parser

The green vegetables supply calcium to the body.	parser wrong in rejecting
*The green vegetables supply calcium strong bones	parser correct in rejecting

Parsing with semantic interpretation

Dictionary with semantic information

dishwasher N
- dishwasher 1
  - *Definition:* A person washing dishes
  - *Category:* Human (therefore animate)
  - *French:* plongeur
  - *German:* Tellerwäscher
- dishwasher 2
  - *Definition:* A machine washing dishes
  - *Category:* Machine (therefore inanimate)
  - *French:* lave-vaisselle
  - *German:* Spülmaschine

jolly ADJ
- *Definition:* Full of merriment and good spirit; fun-loving
- *Modifies:* Human

laughs V
- *Takes subject:* Animate
- *Takes object:*

white ADJ
- *French:* blanc
- *German:* weiss
- white 1
  - *Definition:* A color produced by mixing all rainbow colors, such as in snow.
  - *Modifies:* Non-human (inanimate object or animate object that is not human)
- white 2
  - *Definition:* A race designation used for Caucasian
  - *Modifies:* Human

$_0$ The $_1$ jolly $_2$ dishwasher. $_3$

$_0$ The $_1$ white $_2$ dishwasher $_3$ laughs. $_4$

$_0$ The $_1$ white $_2$ dishwasher $_3$ is $_4$ broken. $_5$

Two traces of semantically augmented parsing

$_0$ The $_1$ jolly $_2$ dishwasher $_3$

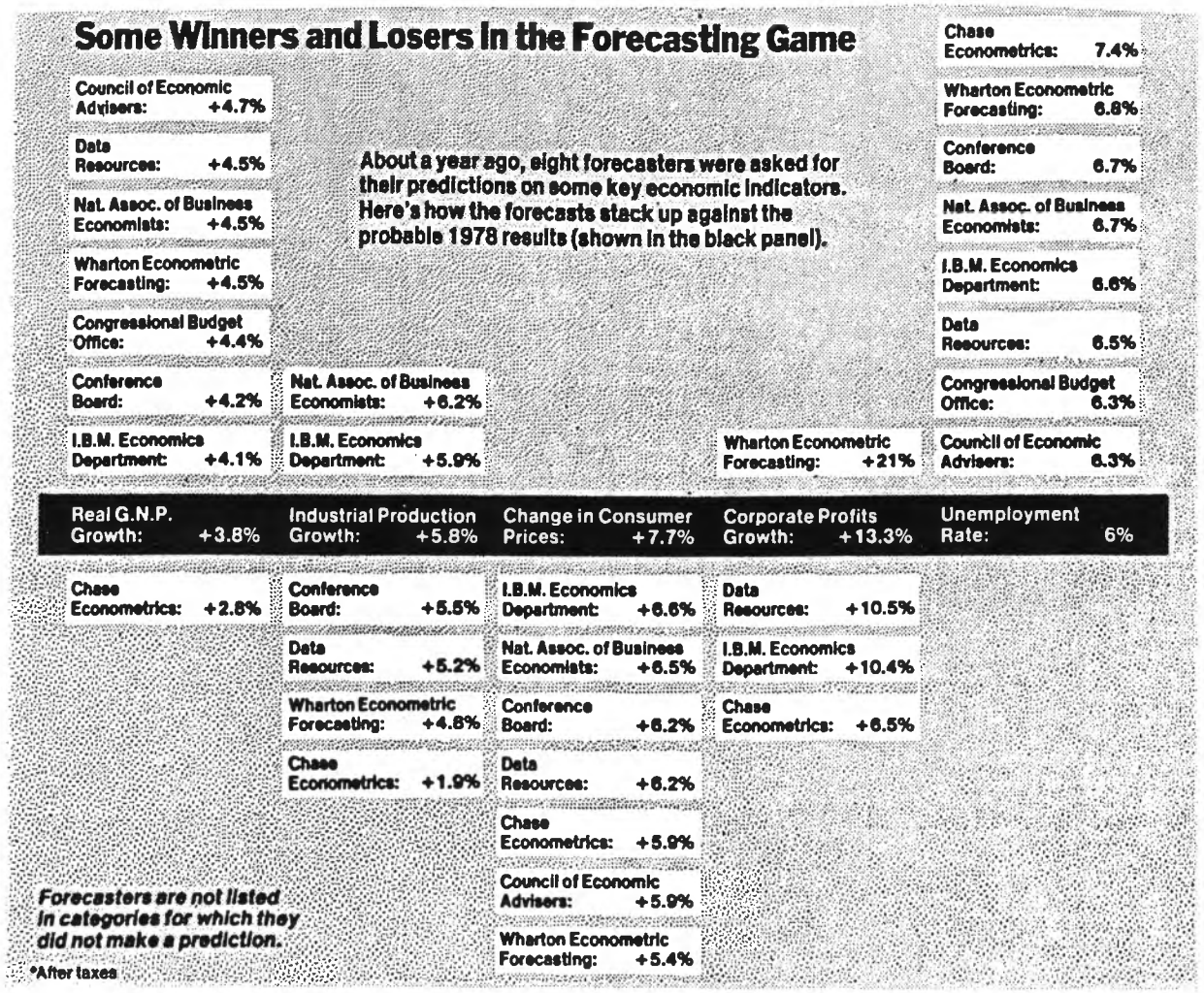
Step	From pos	From state	Segment	To state	To pos
①	0	NP a	the	NP b	1
②	1	NP b	jolly *Requires human noun.*	NP b	2
③	2	NP b	dishwasher *Works only if dishwasher is human* *Select* dishwasher 1	NP c	3

[NP (ART the) (ADJ jolly) (N dishwasher 1)**]**

$_0$ The $_1$ white $_2$ dishwasher $_3$ laughs. $_4$

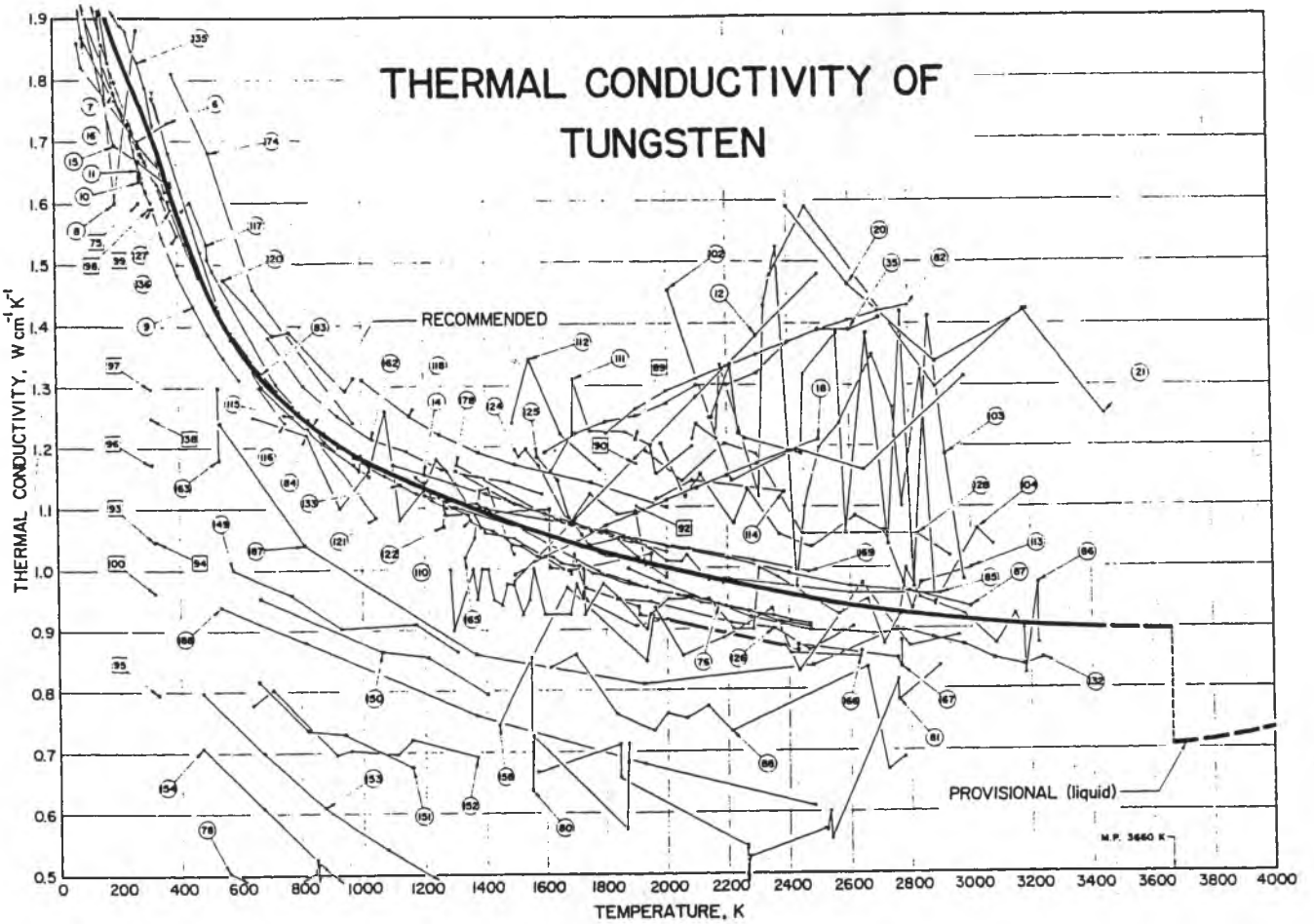
Step	From pos	From state	Segment	To state	To pos
①	**0**	**S^0 a**	**? (consult NP)**	**?**	**?**
②	0	NP^1 a	the	NP^1 b	1
③	1	NP^1 b	white *two meanings:* white 1 *modifies non-human* white 2 *modifies human*	NP^1 b	2
④	2	NP^1 b	dishwasher *two meanings* dishwasher 1 *human* *agrees with* white 2 dishwasher 2 *machine* *agrees with* white 1	NP^1 c	3
⑤	**0**	**S^0 a***	**the white 2 dishwasher 1 *NP1 human*** **the white 1 dishwasher 2 *NP2 machine***	**S^0 b**	**3**
⑥	**3**	**S^0 b**	**? (consult VP)**	**?**	**?**
⑦	3	VP^1 a	laughs *Requires animate subject*	VP^1 b	4
⑧	**3**	**S^0 b**	**laughs *Select NP1***	**S^0 c**	**4**

{S
[NP (ART the) (ADJ white 2) (N dishwasher 1)**]**
[VP (V laughs)**]**
}



New York Times, January 2, 1979, p. D-3

(Tufte 1983, p. 180)



Double-Functioning Labels

Numbers can double-function when used both to name things (like an identification number) and to reflect an ordering. In this graphic (in which the circled numbers fail to double-function), each number identifies a particular study of the thermal conductivity of tungsten, ordered alphabetically by the last name of the first author. If that list were ordered by date of publication instead, then the code would also indicate the time order in which the various conductivity determinations were made. Thus, "1" would indicate the earliest study, and so on; or, alternatively, "61c" would be the third study published in 1961. Such information has interest, since we could see which of the early studies got the right answer. In addition, the movement of the studies toward the "correct" recommended values could be tracked. This extra information requires no additional ink. (Tufte 1983, p. 149 - 150)

CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, *Inspecteur Général des Ponts et Chaussées en retraite.*

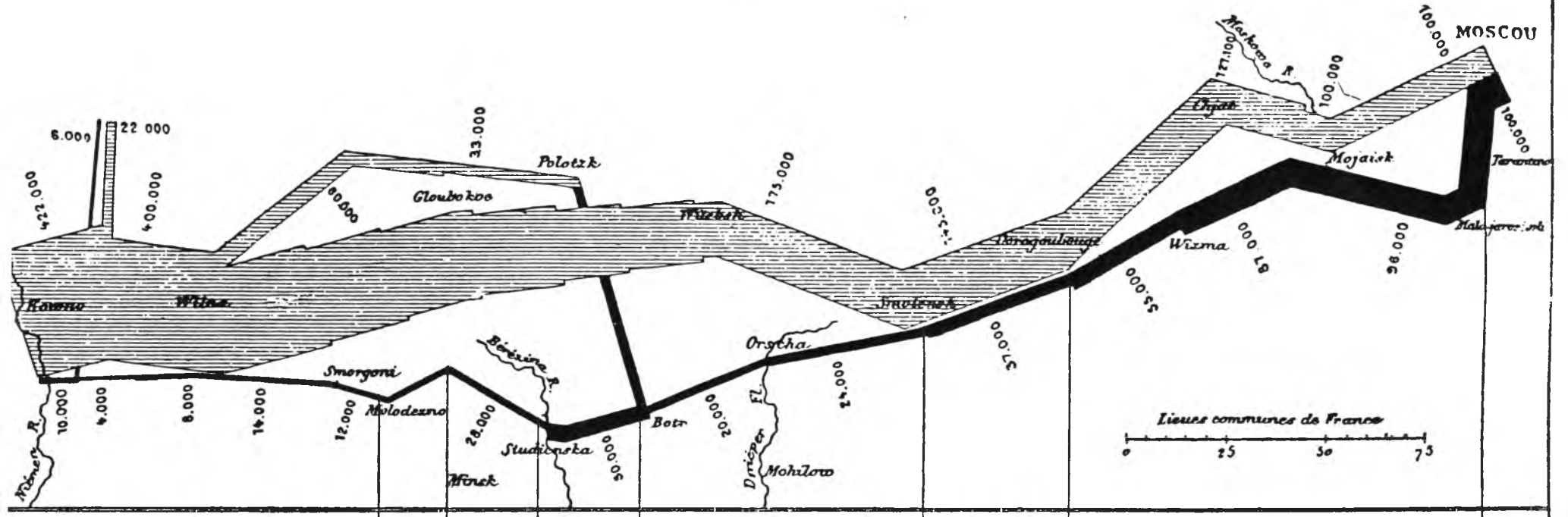
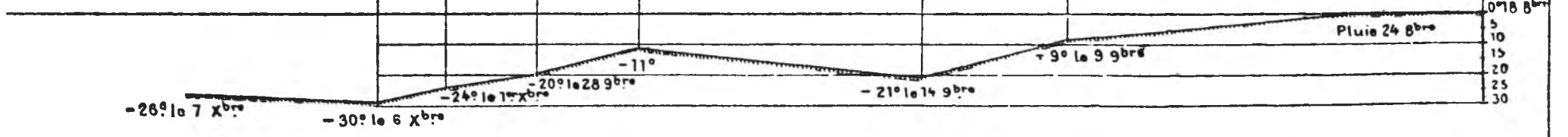


TABLEAU GRAPHIQUE *de la température en degrés du thermomètre de* Réaumur *au dessous de zéro*



Xbre = December 9bre = November 8bre = October

Tufle 1983

Optional.

Example 10

The **Alcohol and Other Drugs Thesaurus (AOD Thesaurus)** provides many examples of meaningful sequence (Wasserman Library, cataloging tools area).

Sample document record from AOD Thesaurus indexing test (next page)

To test the AOD Thesaurus, 20 indexers indexed 25 documents. A cumulative list of the descriptors assigned to each document was then printed. Each descriptor is followed by a list of symbols identifying the indexers who assigned this descriptor. The list is arranged in classified order, facilitating analysis. For example, if the indexers among them assigned several related descriptors, it is easy to see that most indexers covered the basic concept but chose slightly different descriptors; then one can select the best descriptor from those assigned by the various indexers. See the bolded groups at JP8 treatment and MO24.2 public policy on AOD for an illustration. With an alphabetic arrangement of descriptors, this analysis would be much more difficult.

Legend:

C*orrect*
B*road* (assigned descriptor is too broad, above the correct descriptor)
N*arrow* (assigned descriptor is too narrow, below the correct descriptor)
R*elated* (assigned descriptor is related to the correct descriptor)
E*xhaustive* (minor point in document)
T*hesaurus problem* (for example, missing scope note)
W*rong*

CTRL002 Substance-abusing chronically mentally ill client: Prevalence, assessment, treatment, and policy concerns

The bolded groups show assignment of related terms by different indexers.

AB	AODD (ARG) ***B**road*
AB2	AOD abuse (CSRJ, CSRT, CSRP, MAR, BCP) ***B**road*
AM	prevention, diagnosis, and treatment of AODU (CSRJ) ***T**hesaurus problem*
BA	AOD substances of abuse (CSRJ) ***C**orrect*
EC10.10	alcohol interactions (CSRD) ***E**xhaustive*
EC10.8	adverse drug interaction (CSRD) ***E**xhaustive*
FV20.8	assessment (CSRP, BCP) ***B**road*
GA2.12.4	mental dysfunction (BCP) ***N**arrow*
GA2.14.6	dual diagnosis (CSRJ, CSRK, CSRS, CSRA, CSRT, CSRD, CSRP, SHS, MAR, BCP, ...) ***C**orrect*
GA6.10.4.4	chronic disease (CSRJ) ***C**orrect*
GD4	alcohol use disorder (CSRD) ***N**arrow*
GD4.2	alcohol abuse (CSRJ) ***N**arrow*
GD6.2	alcohol related mental disorders (CSRJ) ***N**arrow*
GE2	other drug use disorder (CSRD) ***N**arrow*
GE2.2	other drug abuse (CSRJ) ***N**arrow*
GE4.2	other drug related mental disorders (CSRJ) ***N**arrow*
GY	behavioral and mental disorders (CSRJ, CSRP, MAR) ***N**arrow*
GY2.2.6	other chronic organic psychotic conditions (RIA) ***N**arrow*
HA	screening and diagnostic methods (ARFL) ***B**road*
HB	AODU screening, identification, and diagnostic methods (CSRD) ***C**orrect*
HH2.2	patient AODU history (CAS) ***E**xhaustive*
HK	treatment methods (CSRK, CSRA, CSRD, SHS, ARG) ***C**orrect*
HN10	combined modality therapy (BCP) ***N**arrow*
HX	psychosocial treatment approaches (CSRA, CSRD) ***N**arrow*
HX4.18	cognitive techniques of affect and behavior change (CAS) ***N**arrow*
JK	intervention and treatment (CSRD, ARFL) ***C**orrect*
JM	identification and screening (RIA) ***B**road*
JM2.2	identification and screening for AOD use (SHS, CAS) ***N**arrow*
JP4	patient assessment (CSRJ, CSRK, CSRS, CSRA, CSRD MAR, ARG, CAS, DINF) ***C**orrect*
JP4.4	self report (MAR, RIA) ***N**arrow*
JP8	**treatment** (CSRT, CSRP, BCP, RIA, DINF) ***C**orrect*
JP8.10	**treatment issues** (MAR) ***N**arrow*
JP8.16	**treatment factors** (MAR) ***N**arrow*
JP8.16.2	**patient treatment factors** (CSRK, CSRS) ***N**arrow*
JP8.18.4	**mental health care** (BCP) ***N**arrow*

JT6	mental health services	(BCP) ***E**xhaustive*
JV8	health records	(RIA) ***E**xhaustive*
MO24.2	**public policy on AOD**	(CSRJ, CSRT, BCP, ARFL) ***R**elated*
MO24.2.6	**public policy on other drugs**	(ARFL) ***R**elated*
MO24.2.8	**AOD public policy strategies**	(DINF) ***N**arrow*
MO24.6	**public policy on health**	(CSRK, CSRS, RIA) ***C**orrect*
MT12	employee related issues	(CSRT) ***C**orrect*
NM56	literature review	(CSRK, CSRP, SHS) ***C**orrect*
OF2	alcoholic beverages	(CSRJ) ***E**xhaustive*
PL2.2	incidence and prevalence of AODU	(CSRK, CSRT) ***E**xhaustive*
PL2.6	prevalence	(CSRS, CSRP, SHS, BCP, ARG, DINF) ***E**xhaustive*
PL4	comorbidity	(SHS, ARG) ***E**xhaustive*
PT2.4.6	state wide areas	(CSRK) ***C**orrect*
RB	research and evaluation methods	(MAR) ***B**road*
RC6.2	survey of research	(MAR) ***W**rong*
RM10	assessment of variables and methods	(CSRT) ***C**orrect*
RM10.2	reliability (research methods)	(RIA) ***N**arrow*
RM10.4	validity (research methods)	(RIA) ***N**arrow*
RP	data collection	(CSRD) ***C**orrect*
RP10.6.4	interview	(RIA) ***E**xhaustive*
SG8.2	social work (field)	(CSRP, SHS) ***B**road*
TK4.4.6.2	mentally ill	(CSRK, CSRA, CSRT) ***C**orrect*
TL2	AOD user	(CSRA, CSRD) ***C**orrect*
TT14.2	social worker	(CSRS, DINF) ***C**orrect*

Contents page from Alcohol Research

Lecture 6.1a Supplement

Hypermedia/hypertext ➜ LIS 506 Information Technology

Linear text vs. hypertext	**Typical text is linear in a sequence set by author**: "Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop." Lewis Carroll, Alice in Wonderland, Chapter XII **Hypertext / hypermedia is a collection of text pieces (and images and sound files) with links; the reader can and often must establish her own order through the text** (if indeed the reader goes through the text); this is accomplished by treating the text in blocks (or at least by establishing nodes/locations within the document) and by supplying/permitting links between nodes by which the reader can navigate the text in his or her own order. One could also say that the reader constructs his or her own text. A hypertext can include suggested linear sequences, often indicated by <next> and <previous>.
Major features of hypertext	• fragmented non-linear text form whose components can be rapidly accessed via machine-supported links/relationships under direction of user • interactive • malleable, modular: it is easy to add or revise small pieces • no strong document boundaries (at least in large hypertexts)

Hypertext examples	
World Wide Web	Primary example of hypertext: the **World Wide Web**, in which documents/sites typically have links to other documents/sites; it is the presence of these links that gives the web metaphor. The functionality of hypertext has existed long before, for example, in the form of the research paper with footnotes/bibliography/tables/figures, although WWW makes these links convenient to use.
Wikipedia	
Bible in hypertext format	Links from chapter, group of verses, verse, or word to "Original" version(s); manuscript image(s) Alternative translations Other Bible passages Commentary passages Sermons about the passage (published or own) Entry or subentry in Hebrew/Greek dictionary/grammar Map Archaeological evidence
Fiction examples	"interactive fiction," "Choose your own adventure"

Discussion questions	
	1 How can we design hypermedia systems that support the user in constructing coherent documents?
	2 When should sequence be in the writer's hands, and when should it be in the reader's hands?

Inter-document structures

<table>
<tr><td rowspan="2">Relationships between works (from Dr. Green)</td><td>Continuations and sequels
Answer key
Parodies
Critical reviews
Concordances</td><td>Abstracts
Indexes
Bibliographies
Guides to literature
Translation</td></tr>
<tr><td colspan="2">These are often mentioned in cataloging rules. More examples were presented in the earlier reading Soergel, Integrated information structure interface.</td></tr>
</table>

Citation relationships These are used in a citation index, such as SCI (Science Citation Index) but without differentiating types of citation relationships	• Giving the source of data and ideas in order to enable checking (authenticating), call on an authority, or give credit. • Referring to documents that describe methodology, equipment etc. • Providing background reading; citing whole sections from another document so as to avoid rephrasing an idea already formulated elsewhere but needed for background (avoiding redundancy). • Providing pointers to further reading, including forthcoming work. • Criticizing or correcting previous work (one's own or others).

Notes	1	In hypermedia systems the line between within-document relationships defining the document macrostructure and inter-document relationships becomes blurred.
	2	Citation relationships and relationships (links) in hypermedia systems are often untyped, leaving the reader to guess what the relationship is. In the context of the World Wide Web, there are efforts to allow for the specification of link types.

Lecture 6.1b Supplement

Document example 5: Self assessment memo (p. 190/191)

To: Sue Feldman, CIO
From: Bob Boiko, content management specialist
Subject: Self assessment for year 2000
Date: February 7, 2001
Keywords: Content management; planning; XML; intranet; Web site
URI: www.jasca.com/bboiko/memo20010207-07

Accomplishments in year 2000:

Developed a content management master plan. . . .

Goals for year 2001:

Begin implementation of the content management master plan. . . .

Training needs:

. . .

SGML/XML document type definition (DTD) for self assessment memo

```
<ENTITY % doctype "selfAssessmentMemo" - document type generic identifier >
<!--          ELEMENTS              MIN     CONTENT (EXCEPTIONS)              -->
<!ELEMENT     selfAssessmentMemo    --      (metadata, memoBody)               >
<!ELEMENT     metadata              --      (to, from, subject, date, keywords, URL)>
<!ELEMENT     to                    -O      (#PCDATA)                          >
<!ELEMENT     from                  -O      (#PCDATA)                          >
<!ELEMENT     subject               -O      (#PCDATA)                          >
<!ELEMENT     date                  -O      (#PCDATA)                          >
<!ELEMENT     keywords              -O      (#PCDATA)                          >
<!ELEMENT     URL                   -O      (#PCDATA)                          >
<!ELEMENT     memoBody              -O      (accomplishments, goals  trainingNeeds)>
<!ELEMENT     accomplishments       -O      (#PCDATA)                          >
<!ELEMENT     goals                 -O      (#PCDATA)                          >
<!ELEMENT     trainingNeeds         -O      (#PCDATA)                          >

<!--          ELEMENTS              NAME    VALUE                  DEFAULT-->
<!ATTLIST     selfAssessmentMemo    STATUS  (confidential | public) confidential>
```

A DTD defines a document structure and identifies each element of the structure by a tag. This DTD creates a **selfAssessmentMemo class**. The documents in the memo class must contain two elements, *metadata* and *memBody*. These, in turn, consist of other elements, as listed in ( ). The elements at the bottom of this tree have a data type, in the examples always #PCDATA, which means a character string. ELements can be required or optional; there sequence can be fixed (as in the example) or fixed. This example does not the various syntactic means to specify these options. The memo also has a **status attribute**, whose default value is *confidential*. Alternatively, the status can be *public*.

Another scheme: **O'Neill and Vizine-Goetz** 1989
Note: In the original, they start with *book* and end with *work*.

Work We define a work as a set of related texts with a common source. The term *work* is frequently used inconsistently and, as a result, the distinction between an edition, a printing, and work is often unclear. The term *literary unit* has also been used as a synonym for work. Carpenter found that the words *book* and *work* are used loosely in various definitions and that "sometimes they are even used interchangeably, with a corresponding confusion" (Carpenter, 1981, p. 118).

Using our definition, a work may be composed of substantially different texts. The texts, however, must have been derived either directly or indirectly from a common source. As the text undergoes successive revisions or reexpressions over time, the words and symbols forming later texts may be very different from the original but still represent the same work. In our discussion of text we identified *Moby Dick: La Ballena Blanca and Moby Dick: The White Whale* as separate texts, yet we consider them to be the same work. The translation is closely related to the original and was derived directly from it.

Text [FRBR expression] A text is a set of editions with similar content. The term *text* was introduced by Wilson (1968, p. 6) to describe the content of a book as independent from its physical form. A text is "a sequence of words and auxiliary symbols" which has "no weight and occupies no space" (Wilson, 1968, p. 7). For example, as Hagler and Simmons (1982, p. 74) point out, "the Bantam edition of *Bleak House*, or the 1923 edition, or the Limited edition, may all be identical, word for word, in their textual content, their differences being only in paper, typography, binding, price, and perhaps publisher's name." Thus, a single text comprises three editions. Any edition that has been revised or updated will form a new text. New texts formed by revisions are often identified by numbered edition statements or edition statements such as "New Edition" or "Revised Edition." A new text may also occur as the result of an adaptation or translation. Felix Sutton's abridgement and adaptation of *Ben Hur* for children is a new text. Similarly, *Moby Dick: La Ballena Blanca*, the Spanish translation of *Moby Dick: The White Whale*, is a new text.

Edition [FRBR manifestation] An edition is a set of printings that, at the time of publication, were bibliographically identical. An edition is usually associated with a text. Therefore, if the text changes, so does the edition. However, there are some changes which create a new edition without resulting in a new text. For example, a new edition will be created when a text is republished by a different publisher or with significant changes in type image, or both.

Printing A printing is a set of books by the same publisher which are either printed at one time or printed at different times using the original type image with no more than slight but well-defined variations. As a general rule, the variations permitted within a printing are limited to the correction of minor typographical errors. The books themselves may or may not contain printing information. Commercial publishers commonly display printing information on the verso of the title page. The printing information usually includes the printing number and may also include the printing date.

Book [FRBR item] A book, as defined here, is the bibliographic entity at the lowest level of the hierarchy and is the only one which corresponds to a physical object. All of the other bibliographic entities are abstract concepts. Various terms are used synonymously with *book*, and the term *book is* often used in ways incompatible with our definition. For instance, *item*, *bibliographic item, copy, volume,* and *document* as well as other similar terms have been used interchangeably with the term *book.*

It is the individual book that is used to derive the information necessary for cataloging since, for cataloging purposes at least, all of the books constituting a particular printing are assumed to be bibliographically identical. Therefore, any book can be used to determine the bibliographic properties of the printing.

Advanced exercise: Thinking about rules for corporate entry

The following pages give a number of possible rules and examples for those students with a particular interest in cataloging of documents. (These rules will not be on any test in 571.)

Issue A The first question deals with **choice of main entry**.

A work emanating from a corporate body was obviously, in fact, produced by some person or a group of persons (possibly having a chairperson), and this information is sometimes available to the cataloger. Make a rule about when to make the main entry under person and when under corporate body. Make a rule when to make an added entry for corporate body for those works that have person or title as main entry.

Issue B The following questions deal with **form of entry**, whether main or added entry.

Note: B1, B2, B3 are sub-issues of B for which a rule is needed. B1.1 and B1.2 are alternate rules for sub-issue B1.

B1 Form of name for institutions

Consider the result of applying the following alternative rules for dealing with works entered under a corporate body (either main or added entry) in a large catalog or bibliography from the point of view of ease of searching in the catalog. Consult the examples on p. 241 and 250 which illustrate the problems.

Compare Rule B1.1 and Rule B1.2 with respect to how well they accomplish ease of search.

Rule B1.1. Enter publications emanating from an **institution** (i.e. school, church, radio station, art gallery, etc.) under the place where the institution is located, unless the first word after the initial article is a proper noun or proper adjective. In that case, enter the institution under its name with place added if necessary to distinguish it from other institutions of the same name. Enter the publications of societies (clubs, guilds, fraternities, professional groups, etc.) under the society's name.

	Name in document	Form of entry
B1.1-1	*Metropolitan Museum of Art*	New York, N.Y. Metropolitan Museum of Art
B1.1-2	*University of Maryland*	Maryland (State), University
B1.1-3	*Freer Gallery of Art*	Freer Gallery of Art
B1.1-4	*American Medical Association*	American Medical Association
B1.1-5	*Gardening Club of Haynesville*	Gardening Club of Haynesville

Rule B1.2. Enter a publication emanating from a corporate body under the name of the body.

	Name in document	Form of entry
B1.2-1	*Metropolitan Museum of Art*	Metropolitan Museum of Art, New York, N.Y.
B1.2-2	*University of Maryland*	University of Maryland
B1.2-3	*Freer Gallery of Art*	Freer Gallery of Art
B1.2-4	*American Medical Association*	American Medical Association
B1.2-5	*Gardening Club of Haynesville*	Gardening Club of Haynesville

B1a. What rationale can you perceive for each of the above two rules?

B1b. For each rule try to pin-point where the catalogers and, more importantly, the catalog users would have trouble making decisions. What terms in the rules are particularly difficult to define or interpret?

B2 Names of subsidiary corporate bodies

Consider the fact that corporate bodies are frequently subsidiaries or divisions of other corporate bodies, sometimes with names clearly indicating dependency (like "division") and sometimes with independent names, such as National Research Council, a branch of the National Academy of Sciences. Consider the following possible rules from the point of view of ease of search:

Rule B2.1. List all publications of a corporate body under the name of the parent body.

	Name in document	Form of entry
B2.1-1	*Catalog Code Revision Committee of the American Library Association*	American Library Association
B2.1-2	*National Research Council of the National Academy of Science*	National Academy of Sciences

Rule B2.2. List all publications by sub-divisions or subsidiary bodies **indirectly**. That is, as a sub-heading to the parent body.

	Name in document	Form of entry
B2.2-1	*Catalog Code Revision Committee of the American Library Association*	American Library Association. Catalog Code Revision Committee
B2.2-2	*National Research Council of the National Academy of Science*	National Academy of Sciences. National Research Council

Rule B2.3 List all publications of the divisions or subsidiaries of a corporate body under the subsidiary directly.

	Name in document	Form of entry
B2.3-1	*Catalog Code Revision Committee of the American Library Association*	Catalog Code Revision Committee. (American Library Association)
B2.3-2	*National Research Council of the National Academy of Science*	National Research Council

B3 Name changes of corporate bodies

Corporate bodies are prone to change their names or to use different forms of their name on different publications. Consider the following solutions from the point of view of ease of search:

Rule B3.1 Change all entries to the latest name with references from the older forms of the name.

Rule B3.2 Enter all publications under the original name of the body with references from the newer forms of the name.

Rule B3.3 Enter each publication under the name given on the title page with cross references to previous and later forms of the name.

What about the cost of each rule?

B4 Change in form of name due to a change in the rules

B3 is about name changes in the real world. But how the name of a corporate body is entered in a catalog record also depends on the cataloging rules, such as the rules discussed in this exercise. Rules analogous to Rules B3.1 - B3.3 can be made on how to deal with this problem.

Examples illustrating the problems of form for corporate names

KEY C: Name of the Corporate body
L: Location of the corporate body if it is an institution
P: Person associated with the work (for some help with question)
T: Title of the Work

1. C: Freer Gallery of Art
   L: Washington, D.C.
   T: Dictionary Catalog of the Library of the Freer Gallery of Art, Smithsonian Institution.

2. C: Center for Applied Linguistics
   L: Washington, D.C.
   T: Sociolinguistics (papers from a conference sponsored by the Center)

3. C: Freer Gallery of Art
   L: Washington, D.C.
   T: Eugene and Agnes E. Meyer Memorial Exhibition

4. C: University of Washington
   L: Washington state (for a state institution, the location is the state under ALA rules)
   P: Charles L. Grossman and others (authors)
   T: Migration of College and University Students in the United States (Report of contract between the University of Washington and the U.S. Dept. of Education) The University of Washington is the main entry in the University of Maryland catalog.

5. C: Library of the University of Washington
   L: Washington state
   P: Freda Campbell, compiler.
   T: Filing Rules for the Catalogs of the University of Washington Libraries

6. C: University of Washington
   L: Washington state
   T: Men and learning in modern society (Papers delivered at the inauguration of Charles E. Odegard as president of the University of Washington)

7. C: Public Library
   L: Washington, D.C.
   T: Index to "The Rambler" (a local newspaper feature)

8. C: American Library Association
   T: Bulletin of the ALA

9. C: American Library Association and others
   P: C. Sumner Spalding, general editor
   T: Anglo-American Cataloging Rules (North American Text)

10. C: American Library Association
    P: none, or assume issued by president
    T: Annual Conference Summary Report

Entries according to AACR2 rules

XXX This is a work in progress, some items still need to be checked

Rule 21.1B2 deals with whether to make an entry for the corporate body (whether to establish a relationship)

Rule 24 deals with the form of entry(the form of the entity identifier for the corporate body)

	Entry	**AACR2 Rule**
1	Freer Gallery of Art	24.1
2	Center for Applied Linguistics Assuming this is an independent body. If it is part of a university, it would be different. Would need to research this	24.1
3	This one I'm not sure. I found a rule that said an exhibition should be treated as a corporate body if it reoccurs under the same name. So, if this is true for this exhibition, the entry would be: Eugene and Agnes E. Meyer Memorial Exhibition. If not, the entry would be: Freer Gallery of Art. In order for an exhibition to be the main entry, it must first meet the criteria to be considered a corporate body as stated in AACR2 21.1B1: "[For] art exhibitions, treat as corporate bodies only those that recur under the same name (e.g., Biennale di Venezia, Documenta)." If the exhibition is establishable as a corporate body, it may be used as the main entry heading under categories a) and d) of rule 21.1B2 of AACR2. from http://www.stanford.edu/~kteel/guidelines_mainentry.html	21.1B1
4	Grossman, Charles I am assuming this work to not be administrative in nature or the collective thought of the body	
5	University of Washington. Library. I am considering the library to be a subordinate body.	24.6b, 24.13A
6	University of Washington	24.6b
7	Washington, D.C. Public Library should this entry have a "government" designation? I'm not sure how that should be indicated Also, the preferred name for the locality may be District of Columbia(as used in the name of the library on their Web site)	24.18
8	American Library Association	24.1
9	American Library Association. I am considering AACR to be the collective thought of the body	24.1

10	American Library Association	24.1

Lecture 15.2 Final Review Supplement

Final review. Natural language processing (NLP)

Purposes of natural language processing (NLP)

- Preparing a description of the document

  Descriptive cataloging (e.g. from optically scanned title page)

  Subject indexing

    Multiple index terms

    Assigning a class (from Dewey or LookSmart or Chemical Abstract category), also called document categorization

      Categorizing a document by reading level(more generally: by the audiences for which the document is appropriate)

  Abstracting / summarizing

    Multi-document summaries

- Determining the attitudes, beliefs, or emotions underlying the document (content analysis in sociology and political science or in psychoanalytical methods)
- Determining authorship or other characteristics of the origin of the document
- Preparing a hypertext version of a document, incorporation into a larger hypertext
- Extracting data from a document. Represent the relationships expressed in a document in a more explicit and more easily manipulated way
- Assistance with query formulation
- Natural language interaction with software and systems
- Matching

  Enhanced proximity searching

- Assistance with document creation

  Editing assistance

    Spell check

    Grammar check

  Machine translation, for example on-the-fly translation of Web documents

Natural language answers from databases, text generation

Natural language processing techniques

- Statistical

  Word frequency, phrase frequency, concept frequency. Frequency of words that connote an attitudinal/emotional dimension (content analysis in psychology/sociology/political science). Differential frequency. Looking for the unexpected (such as weighting rare words highly in ranking retrieval results). Association of words with classes / document categories

- Based on text macrostructure - positional approach

  For example: Introduction and conclusions useful source for abstract. Section headings and figure captions useful source for index terms. First and last paragraphs of sections, first and last sentences of paragraphs

- Cue words, phrases, and sentences

  "method", "important result", "new"

- Stemming and other morphological normalization

- Syntactic and semantic analysis

  Parsing of sentences or partial parsing to detect noun phrases

  Parsing with semantic interpretation

  Homonym disambiguation (Subject area of document or Disambiguation rules based on semantic rules (such as *laugh* takes only animate subjects)

  Inter-sentence parsing, resolution of anaphoric references

- Slot filling in frames using parsing or cues

  A technique used equally by human readers and by machine systems.

  Converting natural language statements into entity-relationship expressions. Applying verb case frames. Using cue words to discover type of relationship between two entities, such as *because* or *therefore* indicating causation (See Crombie example in Lecture 6.1).

Knowledge required by NLP

Final review. Precombination vs postcombination

1 Precombination vs postcombination in searching

1.1 Basic problem: Most searches are for topics or themes expressed as compound concepts, such as

the effect of alcohol on the liver or

how to improve test scores of minority children.

In a retrieval system that allows for combining descriptors (as an online system allowing for Boolean query formulations) an index language consisting only of elemental descriptors is sufficient: the user can combine the elemental descriptors that make up her search topic.

But in a retrieval system that allows only searches for single descriptors, as in a card catalog, printed index, shelve arrangement, or a Web subject directory (Yahoo, LookSmart, etc.) the system must provide precombined descriptors for the topics users want to search. The user who wants to find materials on a topic for which there is no precombined descriptor will have difficulty.

Additional reasons for introducing precombined descriptors even in a system that is mainly based on postcombination

1.2 With postcombination, the components of the query formulation may not have the right relationship in the document

Example: A search for

Air transport AND Vehicles

finds a document on

Vehicles used on the metro line to the airport

Need descriptor *Aircraft*

1.3 With postcombination, a combination of elemental descriptors might be ambiguous

Examples:

School	AND	Library	Need descriptors *School library* and *Library School*
Personnel	AND	Administration	Need descriptors *Personnel administration* and *Administrative personnel*

1.4 Requiring the user to combine elemental descriptors may be unnatural

2 Precombination vs postcombination in database organization

Documents are about topics/themes and can be usefully grouped by topics/themes

Examples: Shelf arrangement, Web subject directory, organizing Web search results (often hundreds or thousands of items) into meaningful groups. Need precombined descriptors that define groups (classes).

Problem of arranging precombined descriptors / classes in a meaningful order

Other aspect of same problem: If document can be assigned only one descriptor, that descriptor should express as many of the document concepts as possible; it needs to be precombined.

Relationship top semantic networks; precombined descriptor as an abstraction of what is in common to a group of documents.

3 User problem with systems using a large number of precombined descriptors:

Finding all precombined descriptors under which to search (because of extensive polyhierarchy in a large set of precombined descriptors, one search often requires looking under many, as the DDC and LCC assignments demonstrated)

Solution: descriptor-find index

Assignments Supplement

Assignment 5 Analytical description of an information system

30 min.	**5. Further probe into the structure of DDC**
More DDC structure	Analyze two cases of instructions in the DDC schedules. Discuss each instruction briefly. (1) with respect to combination order (in Case 1, is the order Level of education – Subject or the other way around?) (2) with respect to the effect on the (2.1) exhaustivity and (2.2) specificity of indexing (how many of the aspects for which a document is relevant are represented in the Dewey class (precombined concept) to be used; at what level of specificity); and (3) with respect to the effect on retrieval.

Example

Consider the two books and their Dewey classes

Two facets are coded as follows: Process *Organ system*

(I) **The** action **of Robitussin on the *bronchi***

615.72 Pharmacology and therapeutics > Pharmacodynamics > Drugs affecting the *respiratory system*

(ii) Drug therapy **of *bronchial* disease**

616.23|0 61 Diseases > Dis. of spec. systems and organs > Dis. of the nose, larynx, and access organs > Diseases of the *trachea and bronchi* > Drug therapy

Analysis using the framework given

(1) **Combination order**

(I) books on drug action: Process – *Organ system*

(ii) books on drug therapy: Disease – *Organ system* – Process

(2.1) **Exhaustivity of indexing** – which facets are represented in the class?

(a) books on drug action: Both process and *organ system* are indexed

(b) books on drug therapy: Same

(2.2) **Specificity of indexing** – how specific is the concept from each facet?

(a) books on drug action: Process: Broad *Organ*: Broad (respiratory. system)

(b) books on drug therapy: Process: Broad *Organ*: Specific (trachea and bronchi)

(3) **Effect on retrieval** (recall and discrimination). Consider the query topics

(question 1) Drug action AND *Bronchi*;

(question 2) Drug therapy AND *Bronchi*

q1 cannot be searched specifically, need to broaden *Bronchi* to *Respiratory system*, low discrimination. q2 can be searched specifically.

Advanced example (optional): (question 3) Drug action or therapy AND *Bronchi*

Very tricky: if it is posed as given, it will find documents under 615.23|0 61, but not under 615.72 , giving low recall. Broadening the organ component will increase recall but in this case lower discrimination.

In Case K and Case L you will analyze these problems in the area of education, exploring the effects of the following instruction with 370 Education (slightly edited):

Case K: Level of education versus subject

(a)	**Class special education in a specific subject in 371.9** This takes precedence, (b) and © apply only if not special education
(b)	**Class elementary education in a specific subject in 372.3-372.8.** Example: a book on *Physics experiments for third grade* would be classed under 372.35 Elementary education in specific subjects > Science and technology
©	**Class works on secondary, higher, and adult education in a specific subject with the subject plus the appropriate number under 071 from Table 1 Standard subdivisions** Example: The book *12th grade physics* would be classed under 530|.0712 (530 Physics; 0712 Secondary education, from Table 1) *Freshmen physics* under 530|.0711 (0711 Higher education, from Table 1)

Case K deals with (b) and ©, Case L takes up (a)

K Analyze Case K; write your answers in the proper slots.

(1) Combination order (Level – *Subject* or *Subject* — Level)

(a) books on the elementary level:

(b) secondary or higher level:

(2.1) Exhaustivity of indexing — which facets are represented in the class?

(a) elementary level:

(b) secondary or higher level:

(2.2) Specificity of indexing — how specific is the concept from each facet?

(a) elementary level: Level: *Subject*:

(b) secondary or higher level: Level: *Subject*:

(3) Effect on retrieval (recall and discrimination)

Consider the query topics

Physics AND Elementary school versus

Physics AND Secondary education

Write a very brief analysis

Case L: **Special education, level of education, and subject** (Advanced, optional)

<table>
<tr><th colspan="2">Dewey rules</th></tr>
<tr><td>At 370 Education</td><td>Remember from the instruction with 370 Education:

Class special education in a specific subject in 371.9 Special education.</td></tr>
<tr><td>At 371.9 Special education</td><td>To each subdivision identified by * add (append) the numbers following 371.904 in 371.904 3 - 371.904 7.
At 371.904 4 Programs in specific subjects we are further instructed to add (append) the number following 372 in 372.3 - 372.8

Example 1: the book
High school physics for blind students

would be classed under
371.911| 4|35 Special education > Blind students > in specific subjects/ Science and technology

Class number built following instructions at 371.9, which inherit down to 371.911 as indicated by the *:

4 is from 371.904 4 Programs in specific subjects

35 is from 372.35 [Elementary education in] Science and Technology

Example 2: the book
Accommodations for blind students in high school

would be classed under

371.911| 73

73 is from 371.904 73 Special education at secondary level</td></tr>
<tr><td>Note</td><td>Class 371.9 provides a rich example for this kind of analysis. In Dewey 20 it was possible to specify both educational level and subject within subdivisions of special education.</td></tr>
</table>

L Analyze Case L; write your answers in the proper slots. (Advanced, optional)

(1) Combination order

(2.1) Exhaustivity of indexing — which facets are represented in the class?

(2.2) Specificity of indexing — how specific is the concept from each facet?

:

(3) Effect on retrieval (recall and discrimination)

Invent some query topics for illustration

Write a very brief analysis