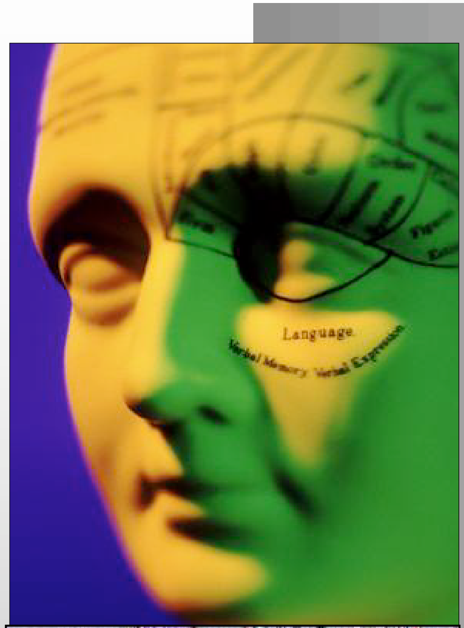
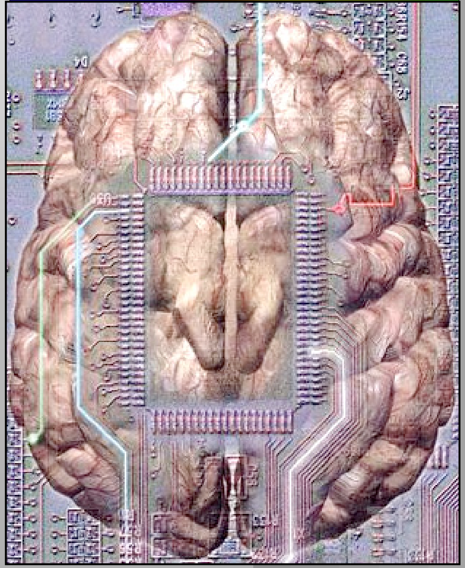
Language Technology I

Hans Uszkoreit

Computational Linguistics at Saarland University
and Language Technology Lab at DFKI

THEORETICAL QUESTIONS





How do people produce and comprehend sentences?

What exactly is the knowledge they must have acquired?

How do they learn languages?

Do we need a full answer before we can apply our knowledge about language and language processing?

INDUSTRIAL CHALLENGE

Bill Gates
Microsoft Chairman and
Chief Software Architect
at IJCAI 2001 in Seattle
August 7, 2001

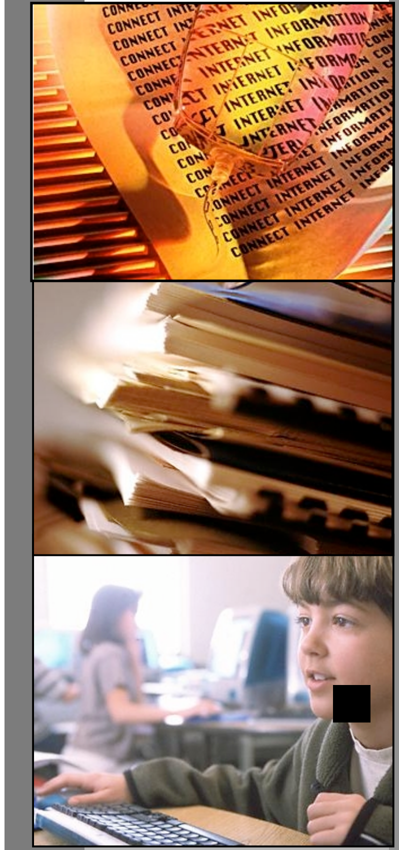
... areas that fit within AI are central to what we're doing, whether it's decision-making learning, language, speech recognition; these are the classic goals of artificial intelligence. We are putting our money where our beliefs are that these things will become real and allow us to build far, far better software products than we have today; and not far better for small audiences. We're talking about software products that many hundreds of millions, if not billions of people will be using and taking advantage of every day...

...

Software can't be so low level that it doesn't understand what the user is trying to do, that it isn't able to look at text and help the user with that.

http://www.microsoft.com/billgates/speeches/2001/08-07aiconference.asp

MAIN APPLICATION AREAS

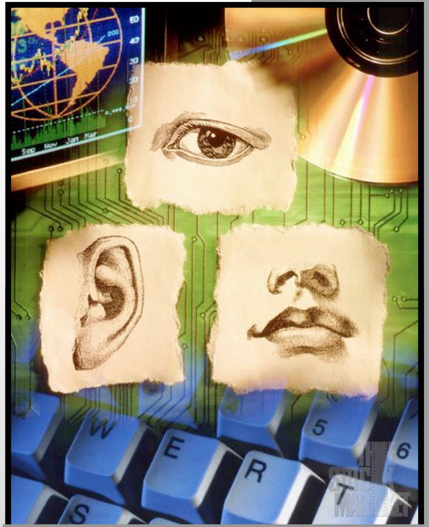


Information and Knowledge Management

Document Authoring, Editing and Publishing

Computer Assisted Language Learning

LANGUAGE & WWW



Language is just one medium in the multimedia setting of the web.

Language is connected with pictures, sounds, movies, VR scenes in many ways.

Language will always remain the primary medium for structuring and accessing all types of information.

→ **Language is the fabric of the web since language is the fabric of knowledge**

INFORMATION MANAGEMENT

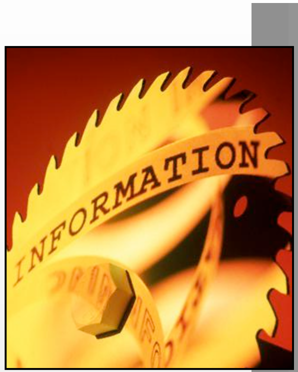
The task of today's

information management

is to gather, maintain and supply

large volumes of digital information.

INFORMATION MANAGEMENT



To this end digital information needs to be…

- created or collected (e.g., gathering, scanning in)
- categorized and indexed (e.g., full-text indexing, classification)
- filtered or ranked (e.g., relevance ranking)
- condensed (e.g., summarisation, information extraction)
- structured (e.g., enriched by hyperlinks, ordered by ontologies)
- delivered (e.g., integrated into intranets, push services)
- presented (e.g., information visualisation)

Tasks of IM

acquisition **(gathering)**

categorisation **(sorting w.r.t topics)**

indexing **(by strings, words, terms, concepts)**

summarisation **(condensing the information)**

information extraction **(relevant data in text)**

translation **(indicative translations)**

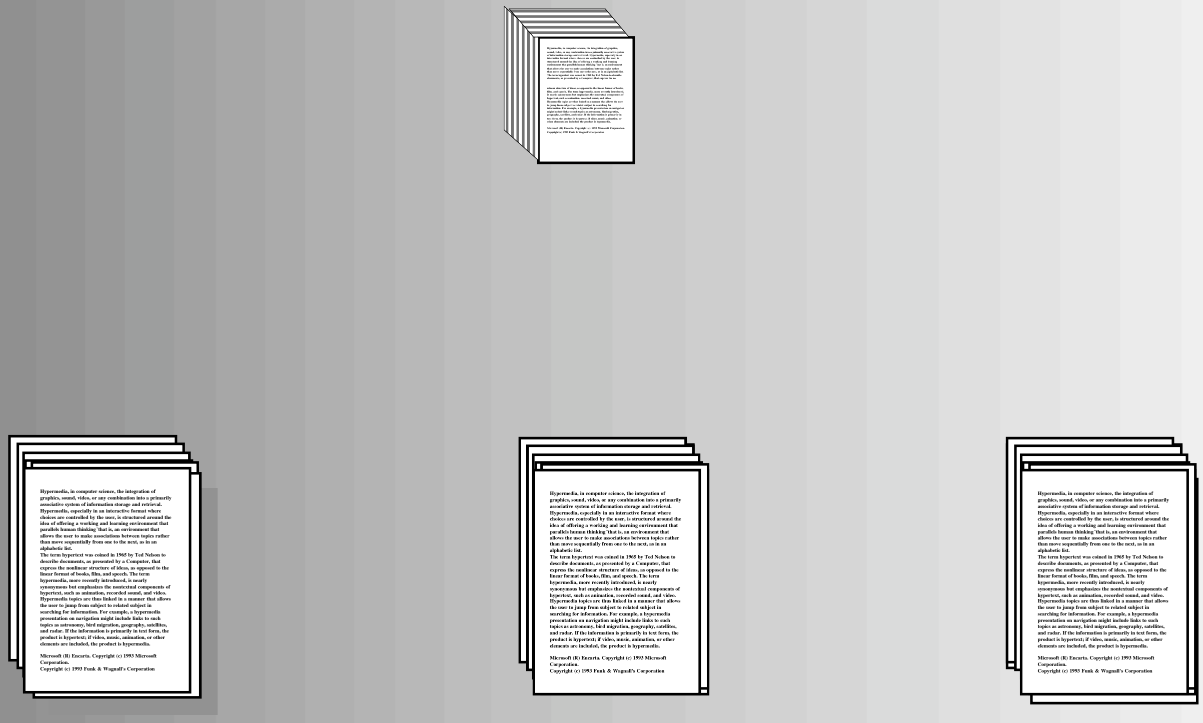
delivering **(filtering, routing, push services)**

presentation **(ranking, structuring, visualising)**

ACQUISITION

- **Scanning**
- **Collecting by Email and Push Services**
- **Gathering from the Net**
- **Sound Recordings**

CATEGORIZATION



Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˜that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.
The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation

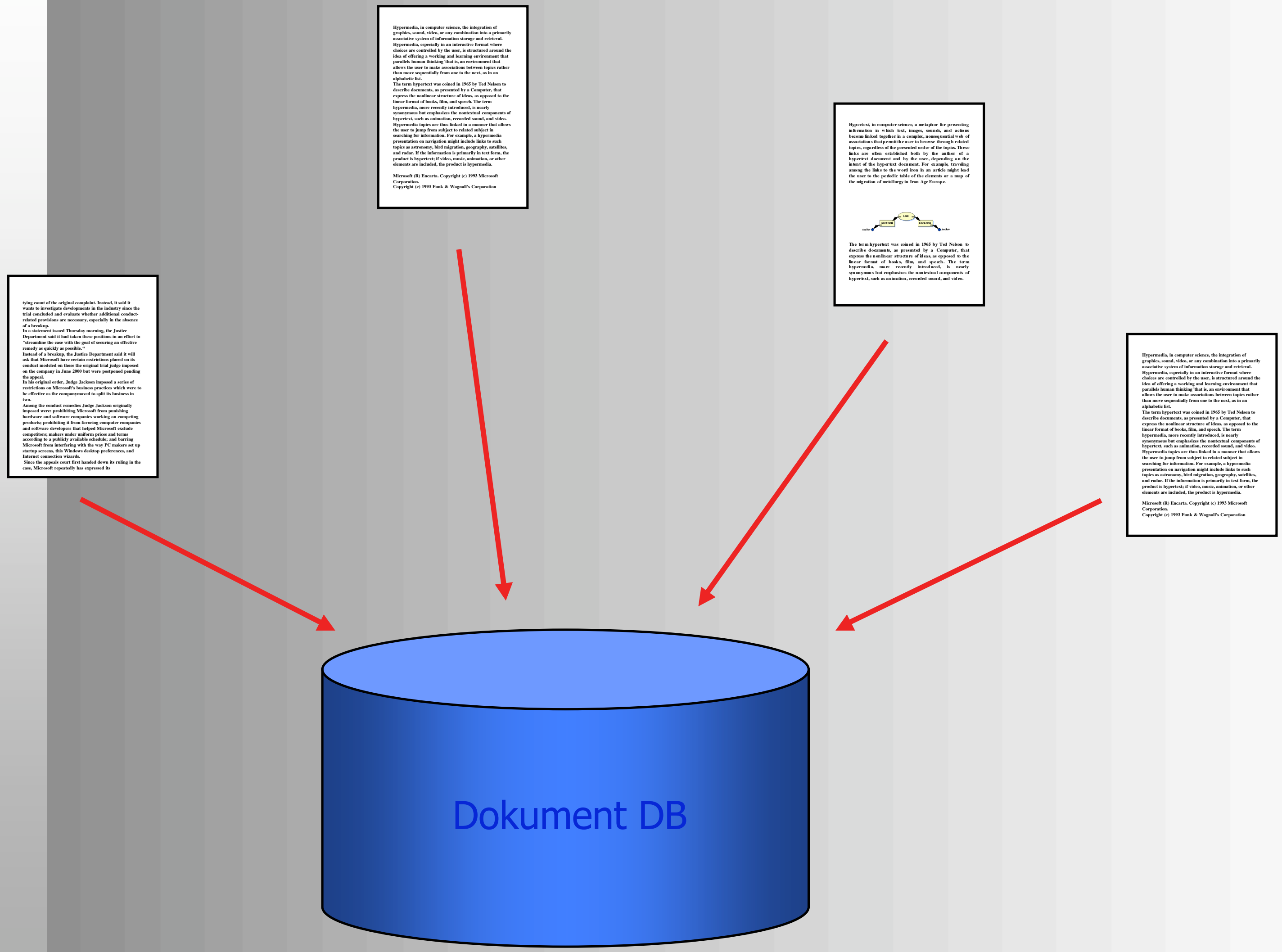
Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˜that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.
The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation

Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˜that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.
The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation

INDEXING



tying count of the original complaint. Instead, it said it wants to investigate developments in the industry since the trial concluded and evaluate whether additional conduct-related provisions are necessary, especially in the absence of a breakup.

In a statement issued Thursday morning, the Justice Department said it had taken these positions in an effort to "streamline the case with the goal of securing an effective remedy as quickly as possible.“

Instead of a breakup, the Justice Department said it will ask that Microsoft have certain restrictions placed on its conduct modeled on those the original trial judge imposed on the company in June 2000 but were postponed pending the appeal.

In his original order, Judge Jackson imposed a series of restrictions on Microsoft's business practices which were to be effective as the companymoved to split its business in two.

Among the conduct remedies Judge Jackson originally imposed were: prohibiting Microsoft from punishing hardware and software companies working on competing products; prohibiting it from favoring computer companies and software developers that helped Microsoft exclude competitors; makers under uniform prices and terms according to a publicly available schedule; and barring Microsoft from interfering with the way PC makers set up startup screens, this Windows desktop preferences, and Internet connection wizards.

Since the appeals court first handed down its ruling in the case, Microsoft repeatedly has expressed its

Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˘that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.

The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation

Hypertext, in computer science, a metaphor for presenting information in which text, images, sounds, and actions become linked together in a complex, nonsequential web of associations that permit the user to browse through related topics, regardless of the presented order of the topics. These links are often established both by the author of a hypertext document and by the user, depending on the intent of the hypertext document. For example, traveling among the links to the word iron in an article might lead the user to the periodic table of the elements or a map of the migration of metallurgy in Iron Age Europe.

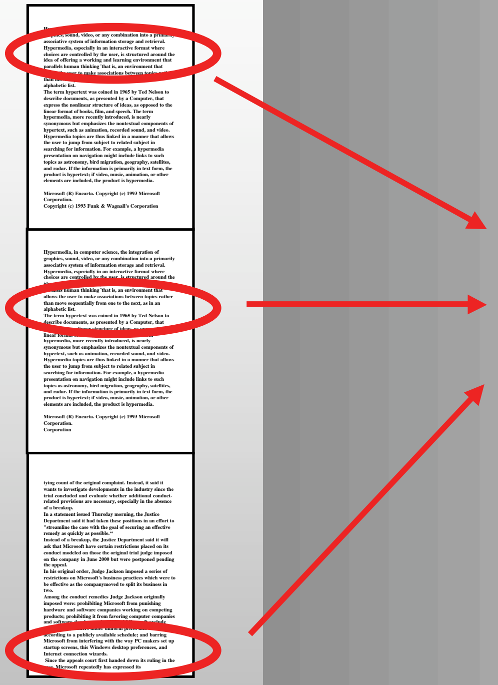
The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video.

Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˘that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.

The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation

SUMMARIZATION



September 6, 2001: 4:39 p.m. ET

[redacted] up:...

nent said Thursday it will not ask that
vo...
ls for the District of Columbia in late
June had overturned a lower court's order....
... it upheld the lower court's conclusion that Microsoft has a monopoly in the market for computer operating systems and maintains that monopoly power

INFORMATION EXTRACTION

- proper names: persons, companies, places...
- special expressions: dates, prices, percentages
- simple relations: company - location, product - price
- complex relations: accident affected parties
  cause
  time
  place
  damage
- answers to questions: Where is the headquarter of IBM?

INFORMATION EXTRACTION

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwaltungssoftware Store Age vorgestellt...
Die neue Version ermöglicht jetzt auch ...

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

...

INFORMATION EXTRACTION

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwaltungssoftware Store Age vorgestellt...
Die neue Version ermöglicht jetzt auch ...

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

...

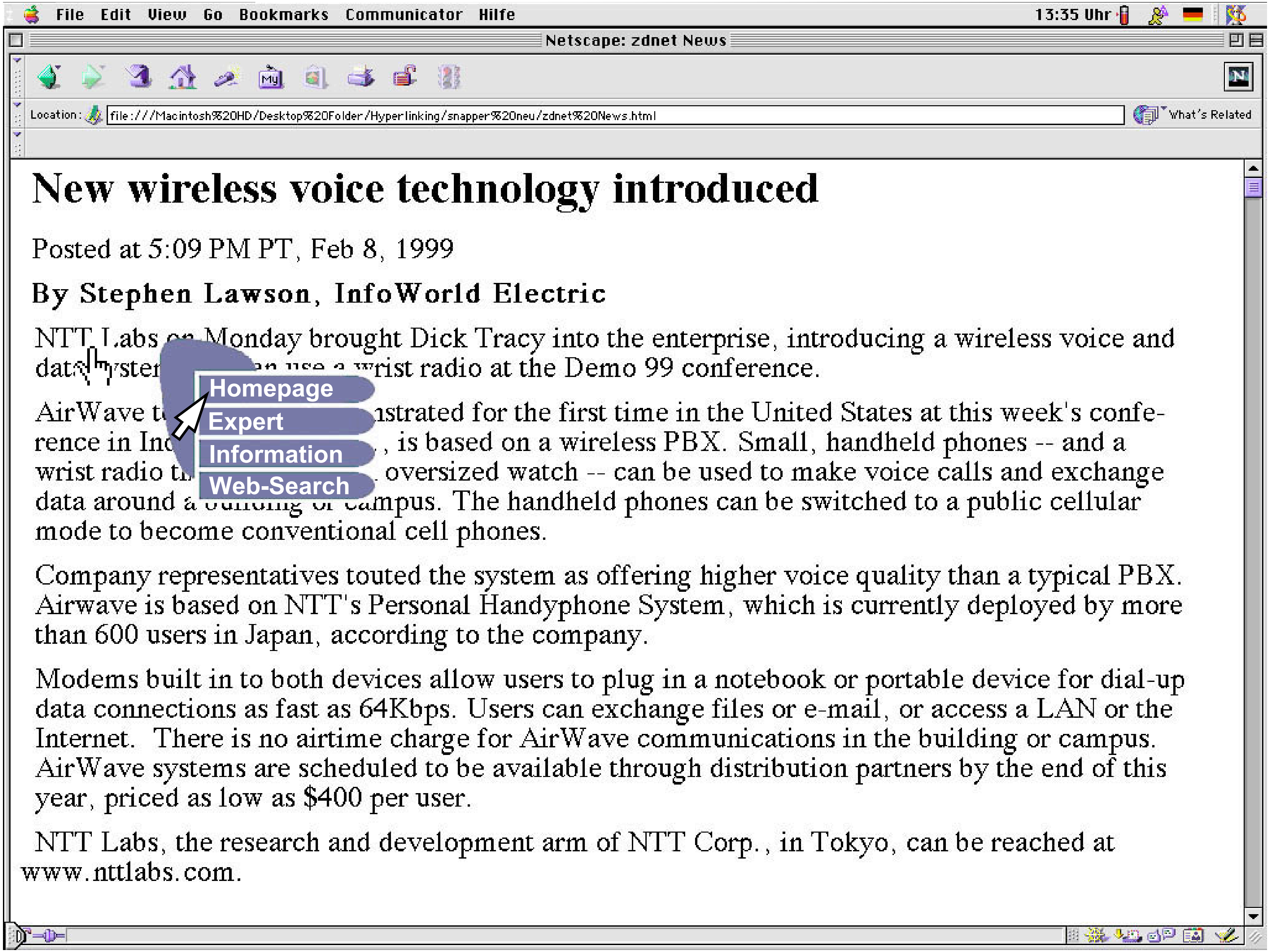
IE RESULT

Firma	96Q4	1996	97Q1	97Q2	97Q3	97Q4	1997	Diff
ComSoft		120Mio					110Mio	-10 Mio
Trade Consult				30 Mio	42,5Mio			12,5 Mio
Z&M					71,0Mio			

GLOBAL INFOSTRUCTURE

A development in three stages

- **Linking machines (ARPANET, INTERNET)**
- **Linking information (today´s WWW)**
- **Creating a dense contextualized associative information network**



New wireless voice technology introduced

Posted at 5:09 PM PT, Feb 8, 1999

By Stephen Lawson, InfoWorld Electric

NTT Labs on Monday brought Dick Tracy into the enterprise, introducing a wireless voice and dat ysten an use a wrist radio at the Demo 99 conference.

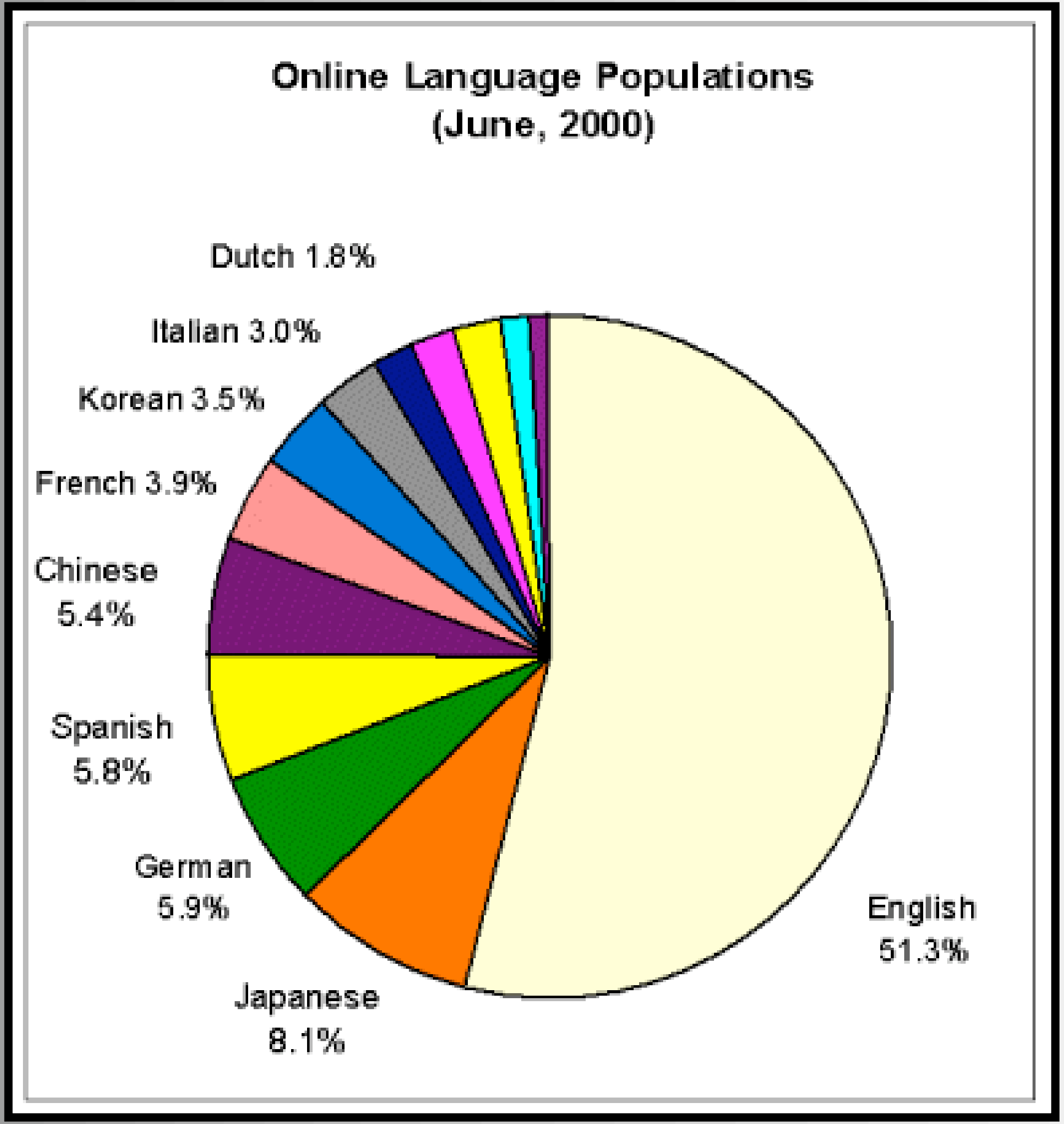
AirWave t nstrated for the first time in the United States at this week's conference in Inc , is based on a wireless PBX. Small, handheld phones -- and a wrist radio th oversized watch -- can be used to make voice calls and exchange data around a building or campus. The handheld phones can be switched to a public cellular mode to become conventional cell phones.

Company representatives touted the system as offering higher voice quality than a typical PBX. Airwave is based on NTT's Personal Handyphone System, which is currently deployed by more than 600 users in Japan, according to the company.

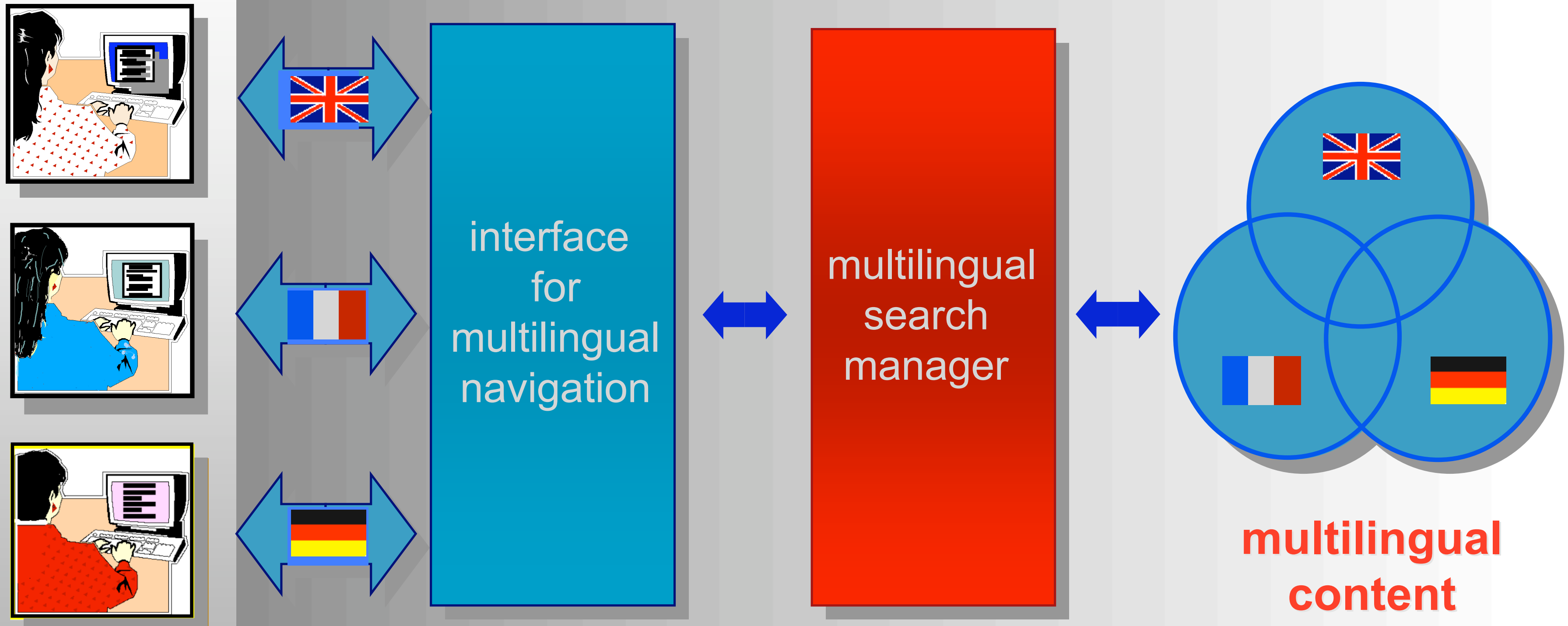
Modems built in to both devices allow users to plug in a notebook or portable device for dial-up data connections as fast as 64Kbps. Users can exchange files or e-mail, or access a LAN or the Internet. There is no airtime charge for AirWave communications in the building or campus. AirWave systems are scheduled to be available through distribution partners by the end of this year, priced as low as $400 per user.

NTT Labs, the research and development arm of NTT Corp., in Tokyo, can be reached at www.nttlabs.com.

MULTILINGUAL WEB



CROSSLINGUAL IR



interface for multilingual navigation

multilingual search manager

multilingual content

The Need

- The majority of participants (86%) is interested in WWW documents written in a known foreign language. Only 22% of our participants are interested in search results in unknown foreign languages.

- Automatic translation of retrieved WWW documents is required by the majority (67%) of the end users.

- 65% of end users want a search engine that translates the query and does the search in the other language.

(Small user survey by Bertelsmann Telemedia)

STATE OF THE ART

95%-98%	Correct recognition of word categories (part-of-speech-tagging)
85%-98%	recognition of names of people, companies, places, products (named-entity-recognition)
95%	statistical recognition of major phrases (HMM chunk parsing)
91%	parsing of newspaper texts by statistically trained parsers (probibilistic context free parsing)
40%-60%	deep parsing of newspaper texts (HPSG or LFG parsing with large lexicon)

DEEP OR SHALLOW ?

Deep Linguistic Processing

accurate
but
brittle

exploits the linguistic knowledge about languages
utilizes grammars and lexicons
derives as much information as possible

versus

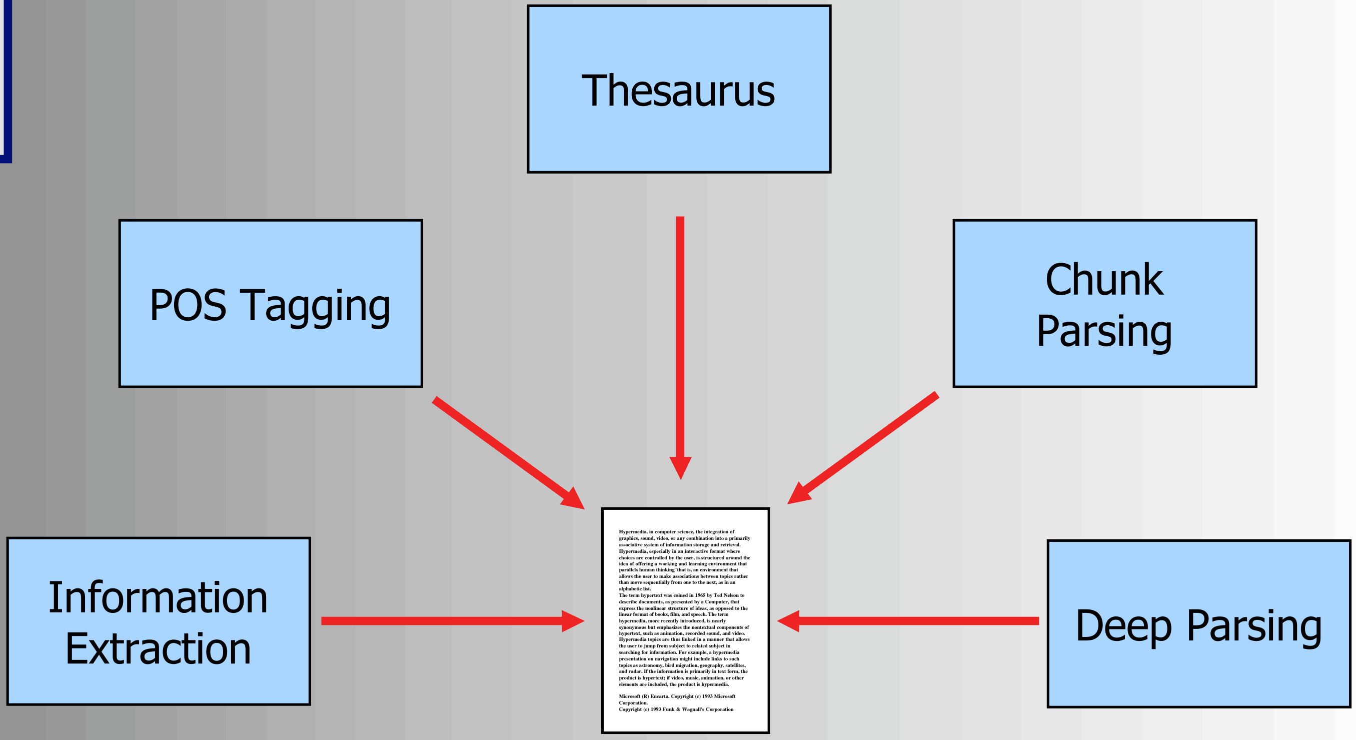
Shallow Linguistic Processing

inaccurate
but
robust

exploits specialized processing methods such as
simple pattern grammars and statistical methods
derives as much information as absolutely needed

COMBINATION

Text Enrichment by XML Tagging



Thesaurus

POS Tagging

Chunk Parsing

Information Extraction

Deep Parsing

Hypermedia, in computer science, the integration of graphics, sound, video, or any combination into a primarily associative system of information storage and retrieval. Hypermedia, especially in an interactive format where choices are controlled by the user, is structured around the idea of offering a working and learning environment that parallels human thinking˘that is, an environment that allows the user to make associations between topics rather than move sequentially from one to the next, as in an alphabetic list.

The term hypertext was coined in 1965 by Ted Nelson to describe documents, as presented by a Computer, that express the nonlinear structure of ideas, as opposed to the linear format of books, film, and speech. The term hypermedia, more recently introduced, is nearly synonymous but emphasizes the nontextual components of hypertext, such as animation, recorded sound, and video. Hypermedia topics are thus linked in a manner that allows the user to jump from subject to related subject in searching for information. For example, a hypermedia presentation on navigation might include links to such topics as astronomy, bird migration, geography, satellites, and radar. If the information is primarily in text form, the product is hypertext; if video, music, animation, or other elements are included, the product is hypermedia.

Microsoft (R) Encarta. Copyright (c) 1993 Microsoft Corporation.
Copyright (c) 1993 Funk & Wagnall's Corporation