Language Technology: An Overview of Commercial Applications and Underlying Technologies

Robert Dale
Centre for Language Technology
Macquarie University
www.ics.mq.edu.au/~rdale

Aims of This Presentation

- To provide a broad awareness of actual and potential Language Technology applications
- To provide a framework for thinking about LT applications in terms of the linguistic resources they need
- To provide a high-level understanding of what's involved in building LT applications

Outcomes

- By the end of this presentation you should:
  - have an understanding of what LT is, and an appreciation of the range of applications that LT enables
  - have an insight into the technologies used in LT applications
  - be able to assess claims about the capabilities of LT applications
  - be aware of the major vendors and suppliers in LT technologies

Presentation Structure

- Part 1: Applications [2 hrs], 12:00–2:00pm
- Lunch Break
- Part 2: Techniques [2 hrs], 3:00–4:00pm

Part 1: Applications

A Definition

- Language Technology involves the application of knowledge about human language in computer-based solutions

Two Drivers for Language Technology

1. The need for intelligent, habitable, natural interfaces:
   - Telephony-based apps need voice capabilities
   - Nobody wants a keyboard on their intelligent microwave
2. The problem of information overload
   - There's too much stuff on the web
   - There's too much stuff in the filing cabinet
   - Nobody has time to read all their email

Related Terms

- Natural Language Processing
- Computational Linguistics
- Speech Technology
- Language Engineering
- Intelligent Text Processing
- Document Processing
- Artificial Intelligence
- Cognitive Science

Two Dimensions

1. Speech versus Text
2. Input versus Output

EXAMPLES	SPEECH	TEXT
INPUT	Desktop dictation packages, speech recognition on the telephone	Optical character recognition, text mining
OUTPUT	Information delivery via synthesised voice	Dynamic web page generation

Principal Components in a Language Technology Application

- Language input
  - recognizing the words
- Language processing
  - reasoning about the words to get at their meaning
- Language output
  - rendering meaning as words

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation
- Dynamic Web Pages

Principal Components in a Language Technology Application

- Language input
  - recognizing the words
- Language processing
  - reasoning about the words to get at their meaning
- Language output
  - rendering meaning as words

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Speech Recognition: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a textual representation of a spoken utterance
- Applications
  - Desktop command and control
  - Dictation
  - Telephony-based transaction and information services
  - Embedded speech
  - Podcast-to-text for search interfaces

Speech Recognition: Fundamental Issues

- Isolated word vs continuous speech
- Vocabulary size
- Speaker dependence vs speaker independence

Speech Recognition: Current State of the Art

- Cheap PC desktop software available: virtually a commodity
- 60–90% accuracy depending on circumstances
- A small number of major players in telephony-based systems
- 'Speech recognition' is almost a commodity; it's what you wrap around it that makes the difference

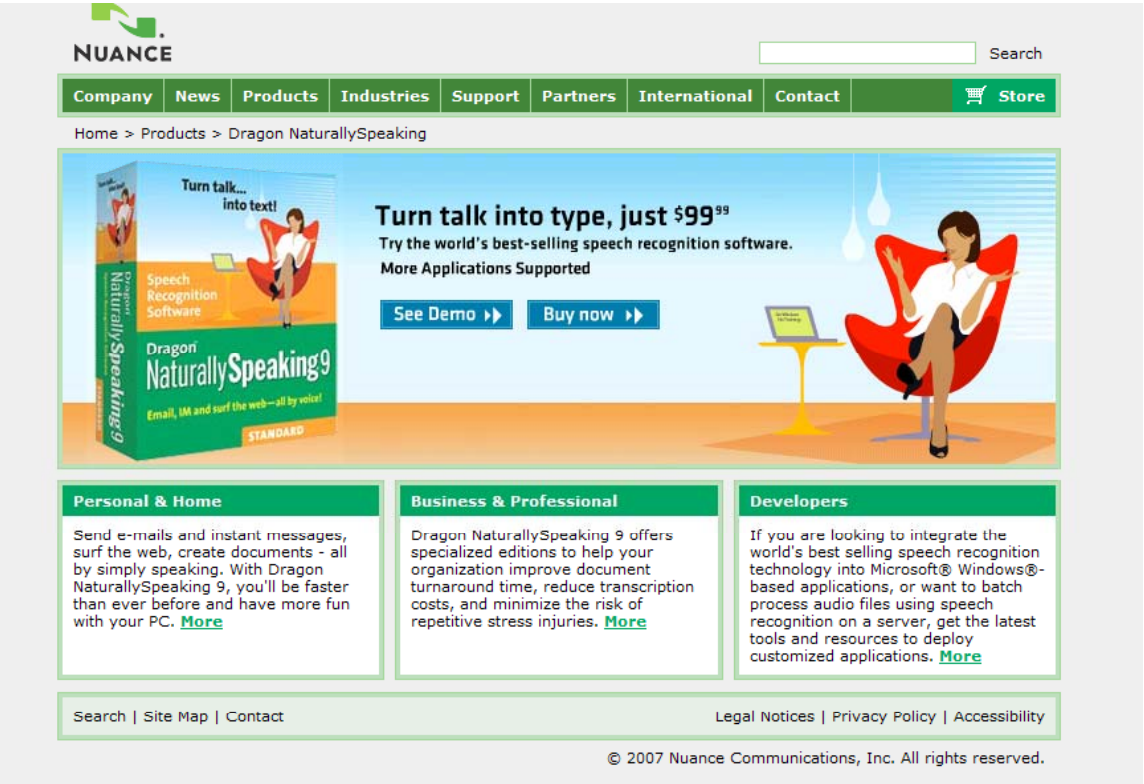
Speech Recognition: Current State of the Art

- Accuracy rates good enough for general dictation and simple transactions, but depends on speaker—your mileage may vary
- Ease of handling errors is important
- Recognition is not understanding!
  - Dictating a letter:
  - Dear George … How about we meet for lunch on Thursday—no, actually—Friday.

Speech Recognition: Fielded Products

IBM ViaVoice: http://www.nuance.com/viavoice/

Speech Recognition: Fielded Products



Dragon Naturally Speaking: http://www.nuance.com/naturallyspeaking/

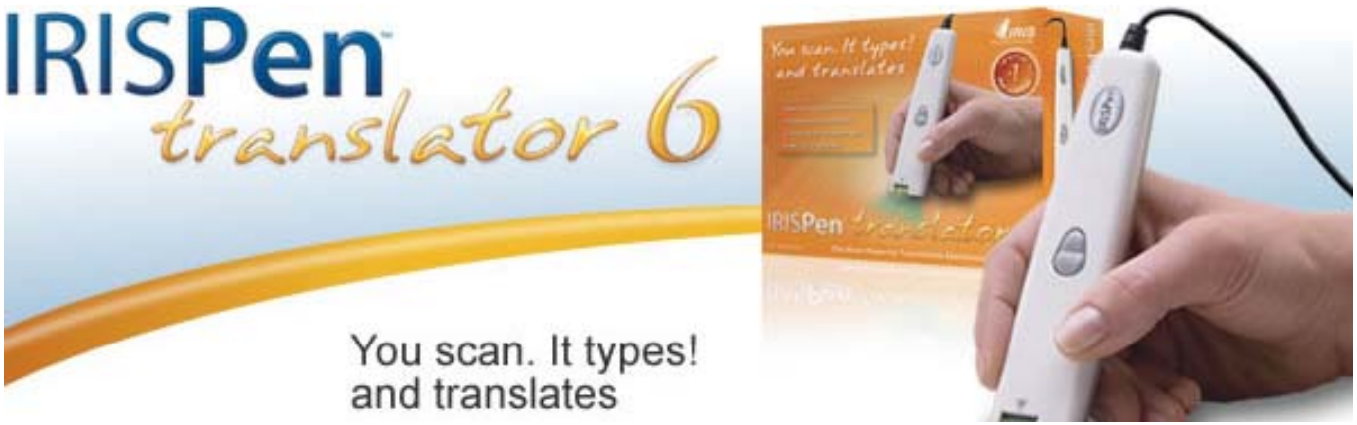
Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Optical Character Recognition: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a computer-readable representation of printed material
- Applications
  - Scanning documents into ASCII form for electronic archival
  - Business card readers
  - Web site construction from printed documents

Menu-Translating Pens



Optical Character Recognition: Fundamental Issues

- Two issues: character segmentation and character recognition
- Problems: unclean data, ambiguity, and new typefaces
- Special fonts aid accuracy (look at your cheque book)
- Many OCR systems use linguistic knowledge to correct recognition errors:
  - N-grams for word choice during processing
  - Spelling correction for post-processing

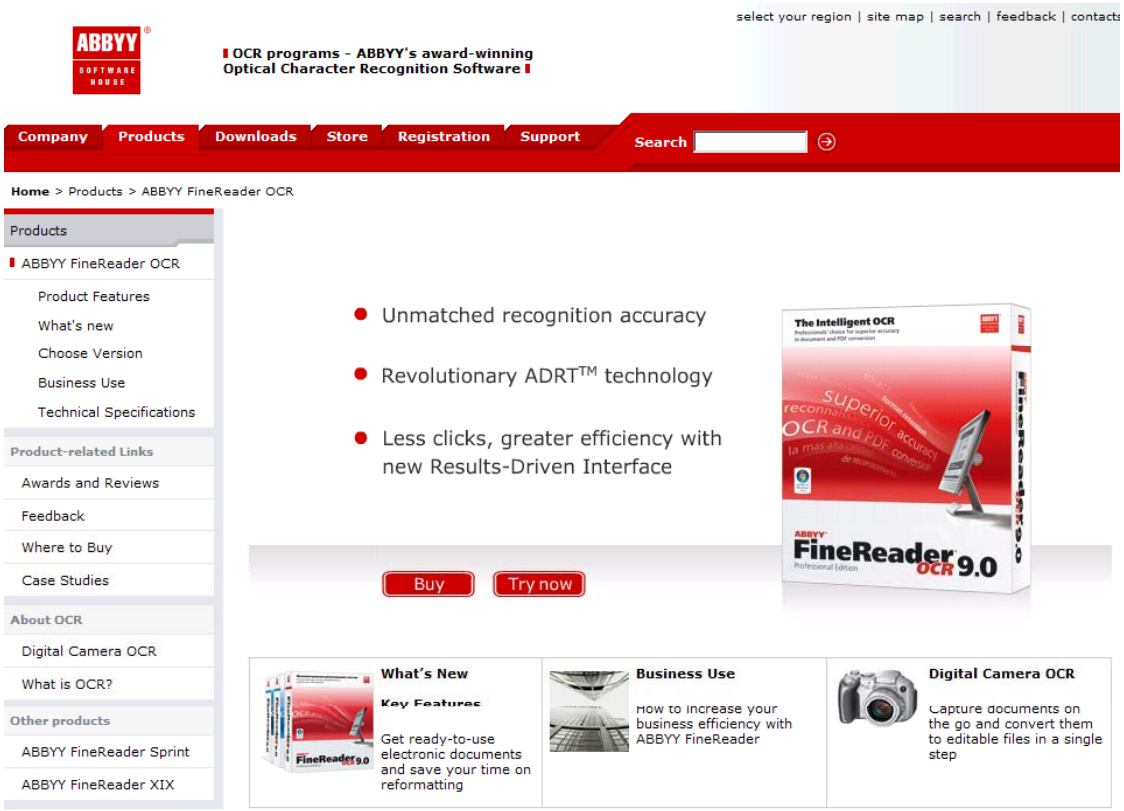
Optical Character Recognition: Current State of the Art

- 90% accuracy or better on clean text
- 100–200 characters per second … as opposed to 3–4 characters per second for typing
- Market development depends on recognising not only characters, but also larger structural elements of documents
- Current apps include TTS ‘read-back’ for proofreading
- Postal services make use of OCR to recognize addresses and assign zip codes in machine-readable form

Optical Character Recognition: Fielded Products

OmniPage: http://www.nuance.com/omnipage/

Optical Character Recognition: Fielded Products



ABBYY FineReader: http://finereader.abbyy.com/

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Handwriting Recognition: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a computer-readable representation of human handwriting
- Applications
  - Forms processing
  - Mail routing
  - PDAs

Handwriting Recognition: Fundamental Issues

- Everyone writes differently!
- Isolated letters vs cursive script
- Better to train the user than to train the system?
  - Apple Newton vs Palm's Graffiti
- Many people can type faster than they can write
  - So, handwriting appropriate where keyboards are not

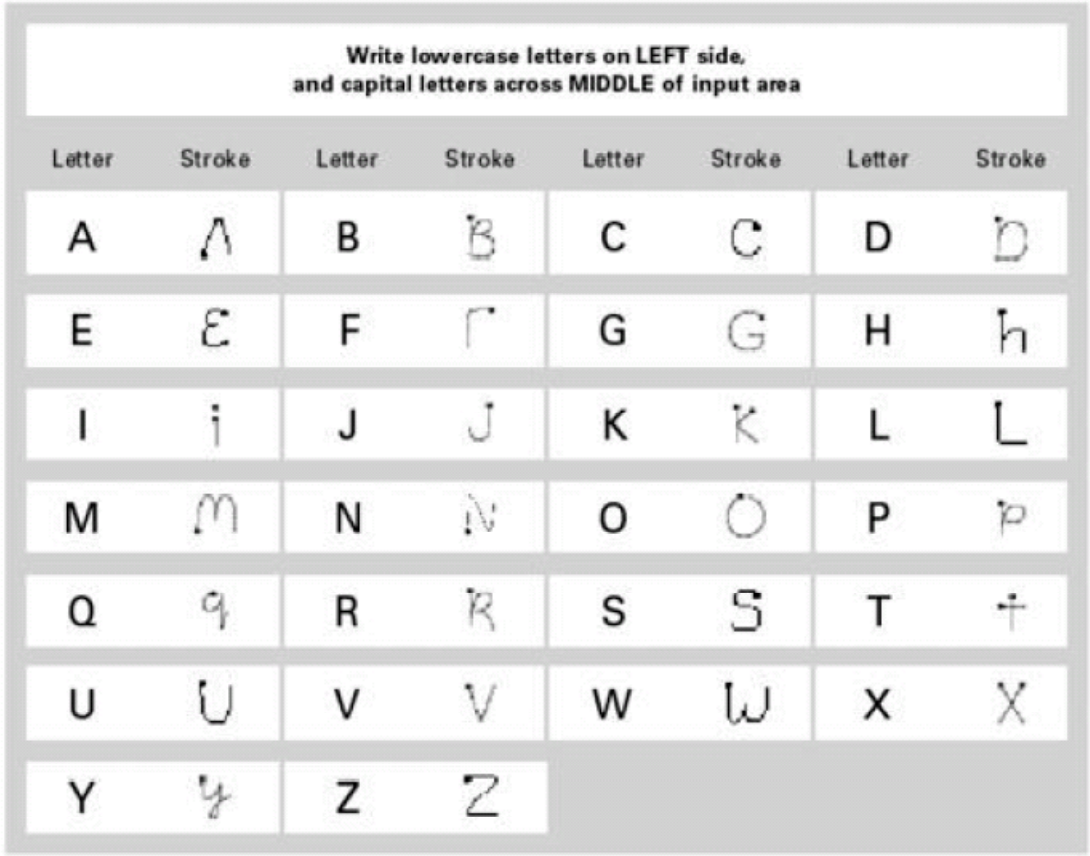
Handwriting Recognition: Current State of the Art

- Generally based on neural network technology
- 5–6% error rate typical for isolated letters
- Good typists tolerate up to 1% error rate on keyboards that generate random errors
- Human subjects make 4–8% errors in isolated character reading, and 1.5% errors given context
- Important distinction between online and offline recognition

Handwriting Recognition: Fielded Products

- Isolated letters
  - Palm's Graffiti (www.palm.com)
  - Computer Intelligence Corporation's Jot (www.cic.com)
- Cursive Script
  - Vision Objects (www.visionobjects.com)
  - Microsoft Tablet PC

Palm's Graffiti: www.palm.com



Principal Components in a Language Technology Application

- Language input
  - recognizing the words
- Language processing
  - reasoning about the words to get at their meaning
- Language output
  - rendering meaning as words

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Spoken Language Dialog Systems: Key Focus and Applications

- Key Focus of the Technology
  - Natural voice interactive dialogs with computer-based systems
  - Spoken dialogue systems communicate with users via automatic speech recognition and text-to-speech interfaces, and mediate the user's access to a back-end database
- Applications
  - Information services: stock quotes, timetables
  - Transaction services: banking, betting, flight reservations

Spoken Language Dialog Systems: Fundamental Issues

- Telephony-based systems cannot afford a training period
- Making a conversation too realistic falsely raises expectations
- Error handling is a significant issue
- Giving dialog initiative to the user increases difficulty

Spoken Language Dialog Systems: Current State of the Art

- Commercial systems operational for limited transaction and information services
  - QTAB betting service
  - American Airlines flight information system
  - Charles Schwab's stock broking system
- Very limited, finite-state notion of dialog
- Limited natural language understanding

Speech Recognition: Fielded Applications

- Speech engine vendors:
  - Nuance
    www.nuance.com
  - Philips
    www.speech.philips.com
- Many third parties building systems around the core speech engines delivered by these companies

Speech Recognition: A Sample Dialog

User calls 8332 8888

System: To book a taxi, say yes. To check a booking, say check. For anything else, say no.

User: Yes.

System: Are you travelling from 14A Spears Avenue Balmain?

User: Yes.

System: And what suburb are you going to?

User: Milson's Point.

System: How many passengers are travelling?

User: One.

System: Are you ready now?

User: Yes.

System: Confirming your booking for a taxi from 14A Spears Avenue Balmain to Milson's Point for one person leaving now. Is that correct?

User: Yes.

System: Your booking for the next available taxi is confirmed. Your reference number is . . .

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Search and Information Retrieval: Key Focus and Applications

- Key Foci of the Technology
  - Concept-based search: moving beyond documents as bags of words
  - Named entity recognition
- Applications
  - Intelligent web search
  - Desktop document retrieval

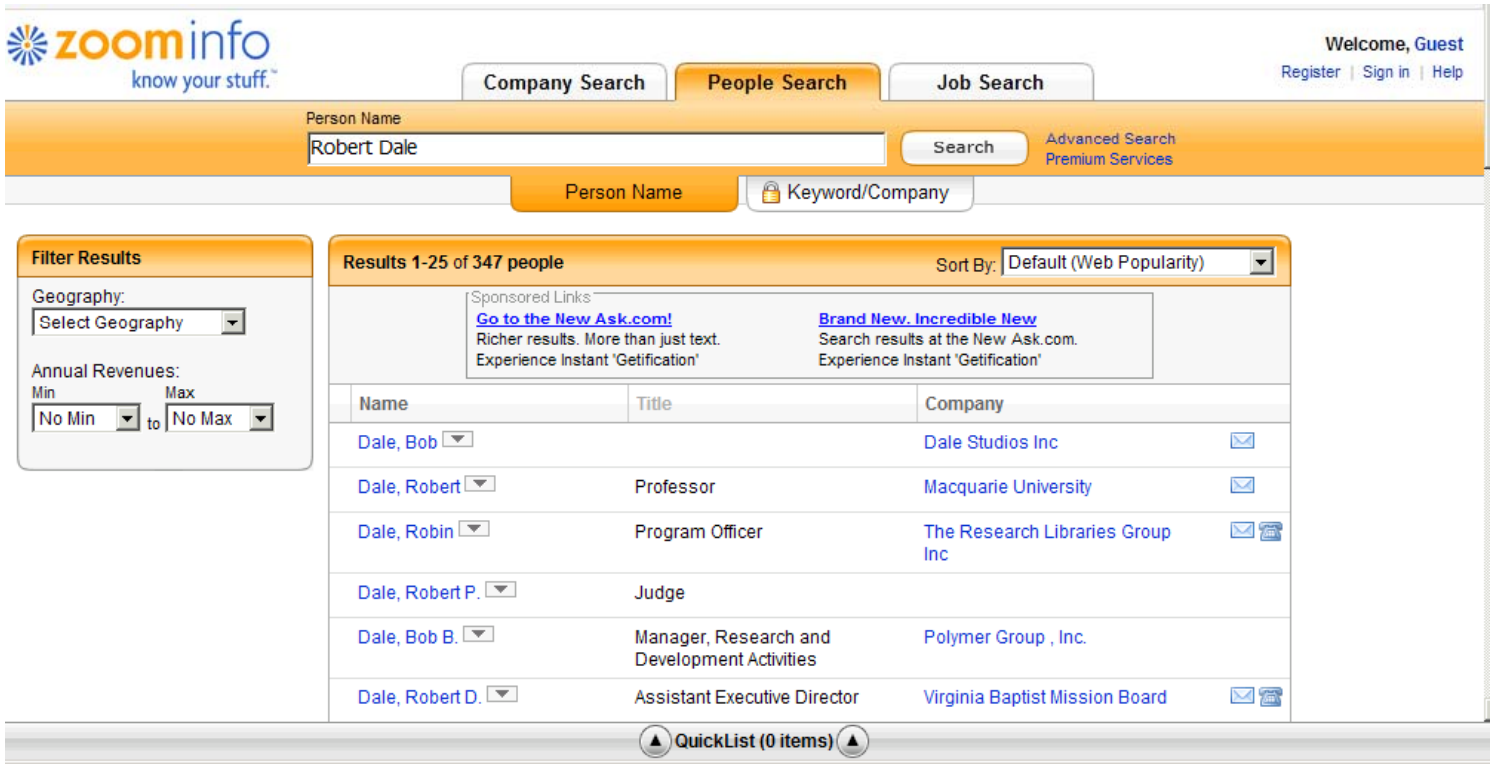
Search and Information Retrieval: Fundamental Issues

- Major failure in IR systems: vocabulary mismatch
  - The information need is described using words other than those used in relevant documents
  - Solved by automatic expansion of the query
- Named Entities:
  - One person or organisation can be referred to by many name variants
  - Many persons or organizations can share the same name

Search and Information Retrieval: Current State of the Art

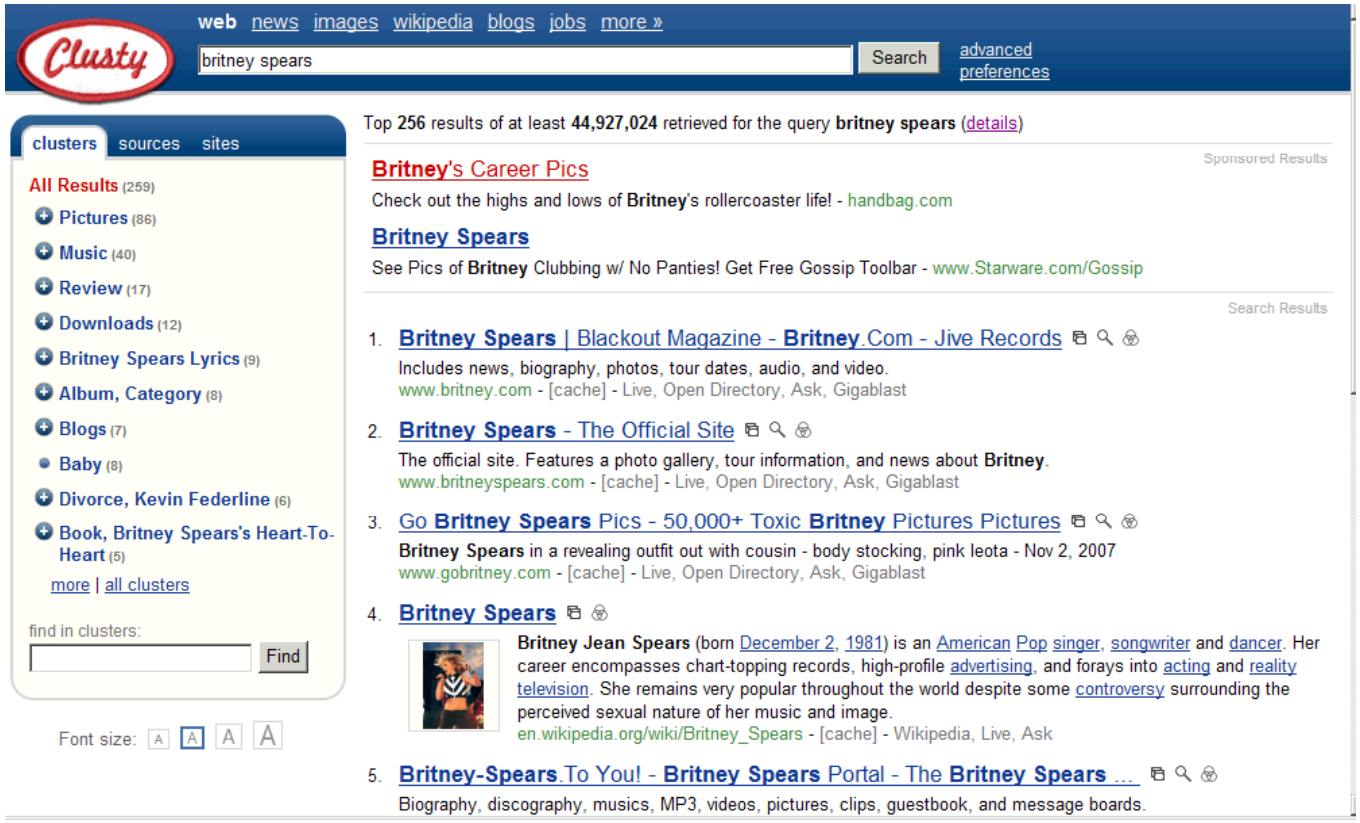
- Google, Yahoo!, Microsoft Live Search …
- Thesaurus-based vocabulary expansion
- 'Did you mean …?'
- Clustered Search
- People Search
- Predominantly rule-based Named Entity Recognition
- Limited linguistic analysis to determine phrases rather than words

Search and Information Retrieval: People Search



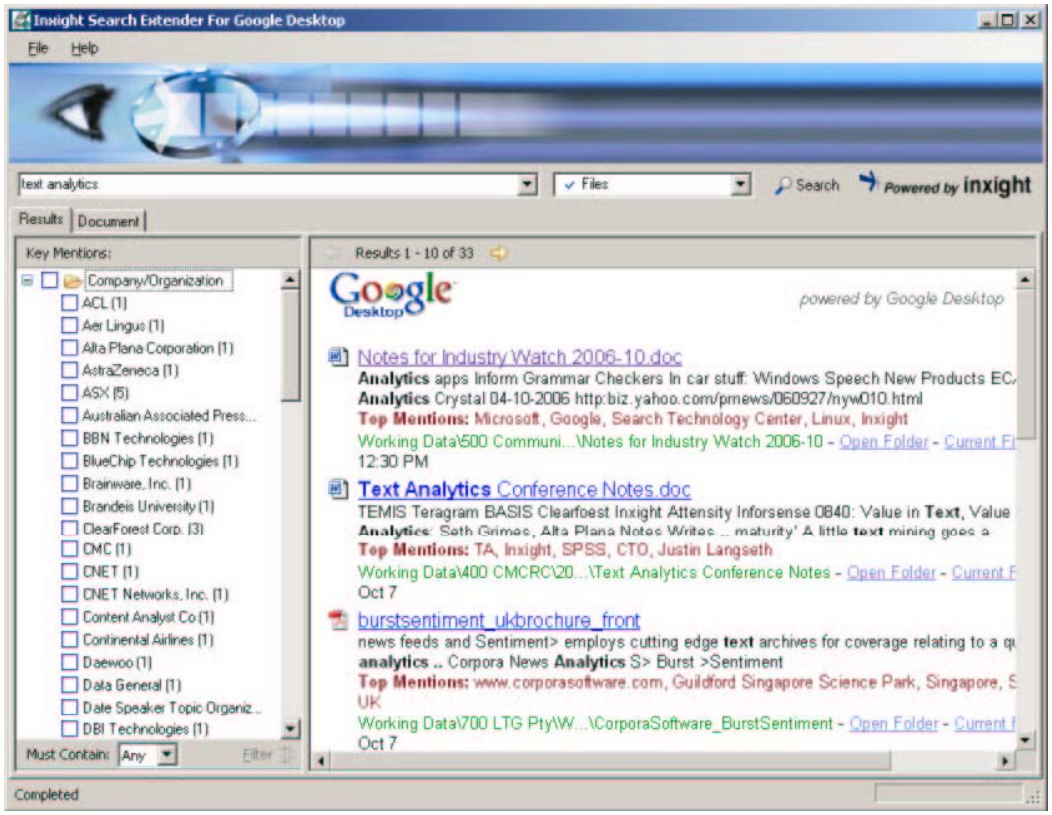
http://www.zoominfo.com/

Search and Information Retrieval: Clustered Search



http://clusty.com/

Search and Information Retrieval: Named Entity Recognition



http://www.inxight.com/products/se_google/

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Grammar and Style Checking: Key Focus and Applications

- Key Foci of the Technology
  - Broad-coverage grammar development
  - Getting beyond syntax to higher-level issues
- Applications
  - Spell checking
  - Grammar checking
  - Style checking

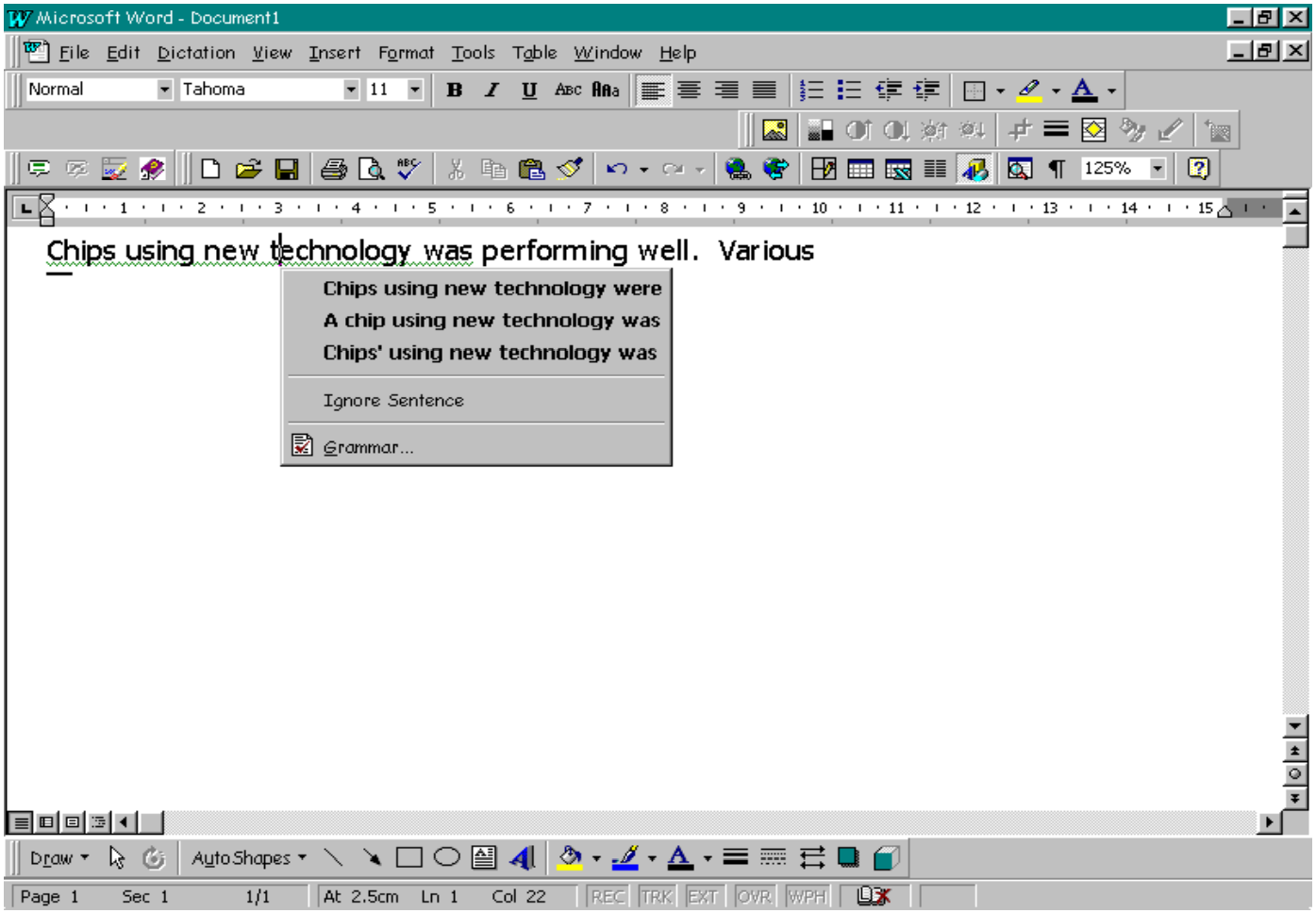
Search and Information Retrieval: Fundamental Issues

- Nobody has a parser that can correctly handle all the correct sentences in the language, so recognizing errors is a bit of a problem …
- The detection of many grammatical errors require an understanding of meaning or communicative intent
- Errors in grammar are only one aspect of writing where assistance is required

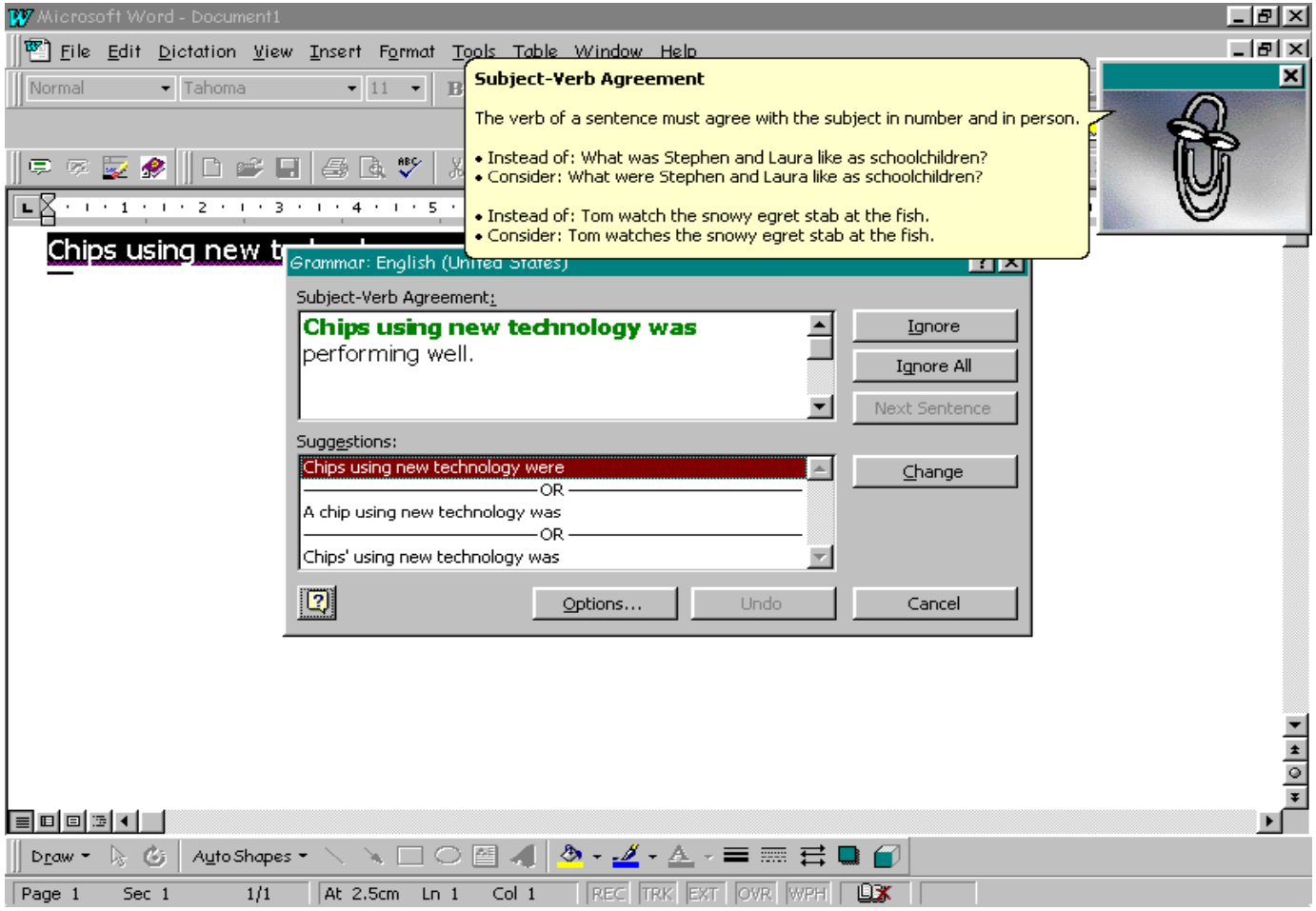
Search and Information Retrieval: Current State of the Art

- Microsoft Word's Grammar Checker
  - Much maligned by academics, it really does provide help for those who need it
  - Still a problem with false positives
  - Still a problem with high expectations

Writing Assistance: Word's Grammar Checker



Writing Assistance: Word's Grammar Checker



Writing Assistance: WhiteSmoke

- US$79.95 up
- www.whitesmoke.com

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Machine Translation: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a version of a document in a language other than that used in the original document
- Applications
  - Desktop and web-based translation services
  - Spoken language translation services

Machine Translation: Bowne Global Solution's iTranslator

- Source English:
  - With worldwide translation and documentation services, L&H can help you do business in any market around the globe. Our technology and top-line people enable you to communicate with employees and customers anywhere.
- Translated into French …:
  - Avec traduction mondiale et la documentation entretient, L&H peut vous aider à travailler dans tout marché autour du globe. Notre technologie et gens de la sommet-ligne vous permettent de communiquer n'importe où avec les employés et les clients.
- … and back into English:
  - With world translation and the documentation maintains, L&H can help you to work in all market around the globe. Our technology and people of the summit-line allow you to communicate where with the employees and the customers.

Machine Translation: Systran's Web-Based Translator

- Source English:
  - With worldwide translation and documentation services, L&H can help you do business in any market around the globe. Our technology and top-line people enable you to communicate with employees and customers anywhere.
- Translated into Greek …:
  - **Με τις παγκόσμιες υπηρεσίες μεταφράσεων και τεκμηρίωσης**, L&H **μπορεί να σας βοηθήσει να κάνετε επιχειρήσεις σε οποιαδήποτε αγορά σε όλη την υδρόγειο. Οι άνθρωποί μας τεχνολογίας και κορυφαίος-γραμμών επιτρέπουν σε σας για να επικοινωνήσουν με τους υπαλλήλους και τους πελάτες οπουδήποτε**.
- … and back into English:
  - With the world services of translations and documentation, L&H it can you help you make enterprises in any market in the all globe. Our persons of technology and leading-line allow in you in order to they communicate with the employees and the customers anywhere.

Machine Translation: Fundamental Issues

- The broad coverage required by mainstream translation technologies exacerbates ambiguity problems
- Effectively limited to literal language use
- Main approaches:
  - Transfer
  - Interlingua
  - Example-based
  - More recently, statistical machine translation
- Real systems often Machine-Assisted Translation

Machine Translation: Current State of the Art

- Broad coverage systems already available via the Web
- Fast turnaround, acceptable error rate for gisting
- Higher accuracy can be achieved by carefully domain-targetted systems
- Controlled languages such as Caterpillar English maximise likelihood of accurate translation

Machine Translation: Fielded Products

- Systran—used by AltaVista
  - http://www.systran.co.uk/
- Language Weaver
  - http://www.languageweaver.com

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Text Summarisation: Key Focus and Applications

- Key Focus of the Technology
  - Producing a version of a document that is shorter than the original document
- Applications
  - Information browsing
  - Voice delivery of web pages and email

Text Summarisation: Fundamental Issues

- There are different kinds of summaries:
  - Informative vs indicative
- Real summarisation requires real understanding
- Quality of 'knowledge-free' summarisation relies on aspects of the document other than content

Text Summarisation: Current State of the Art

- Commercial systems work on a 'sentence-extraction' model
- Sentences extracted on basis of
  - location
  - linguistic cues
  - statistical information
- Relatively knowledge-free but broad coverage as a result

Text Summarisation: Fielded Applications

- InXight's Summarizer (www.inxight.com)
- MS Word's Summarisation Tool
- Xerox demo incorporating OCR, MT and summarisation

MS Word's Summarizer: 10% Summary

SummerFest 2006
SummerFest is the premier event on the HCSNet calendar. Held annually, SummerFest is an inspiring week-long conference designed to bring together participants from the range of disciplines that make up Human Communication Science. SummerFest 2006, which was held from 27 November to 1 December 2006 at the University of Sydney, encompassed our annual two-day Summer School, our innovative speed papers session, and three interdisciplinary workshops. The aim of SummerFest is to bring together as many members of HCSNet as possible, to provide students with opportunities to hear the latest research and trends in the range of disciplines that make up HCSNet, and to allow researchers in HCSNet to meet fellow members, hear of their research interests, and begin to develop collaborations that cross discipline boundaries. SummerFest 2006 attracted over 200 attendees, of whom over half were students from 15 different tertiary institutions across Australia. Other participants included international visitors, academics and faculty staff, and researchers from industry.

MS Word's Summarizer: 25% Summary

SummerFest 2006
SummerFest is the premier event on the HCSNet calendar. Held annually, SummerFest is an inspiring week-long conference designed to bring together participants from the range of disciplines that make up Human Communication Science. SummerFest 2006, which was held from 27 November to 1 December 2006 at the University of Sydney, encompassed our annual two-day Summer School, our innovative speed papers session, and three interdisciplinary workshops. The aim of SummerFest is to bring together as many members of HCSNet as possible, to provide students with opportunities to hear the latest research and trends in the range of disciplines that make up HCSNet, and to allow researchers in HCSNet to meet fellow members, hear of their research interests, and begin to develop collaborations that cross discipline boundaries. SummerFest 2006 attracted over 200 attendees, of whom over half were students from 15 different tertiary institutions across Australia. Other participants included international visitors, academics and faculty staff, and researchers from industry.

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Search and Information Retrieval
- Writing Assistance
- Machine Translation
- Text Summarisation
- Question-answering Systems

Question-Answering Systems: Key Focus and Applications

- Key Focus of the Technology
  - Given a natural language query, produce an appropriate response
- Applications
  - Web-based information services
  - Desktop help systems

Question-Answering Systems: Fundamental Issues

- Limiting coverage to short questions provides some restriction on syntactic structure but leaves open vocabulary issues
- Real questions often contain presuppositions and contextual assumptions:
  - Where can I find my class timetable?

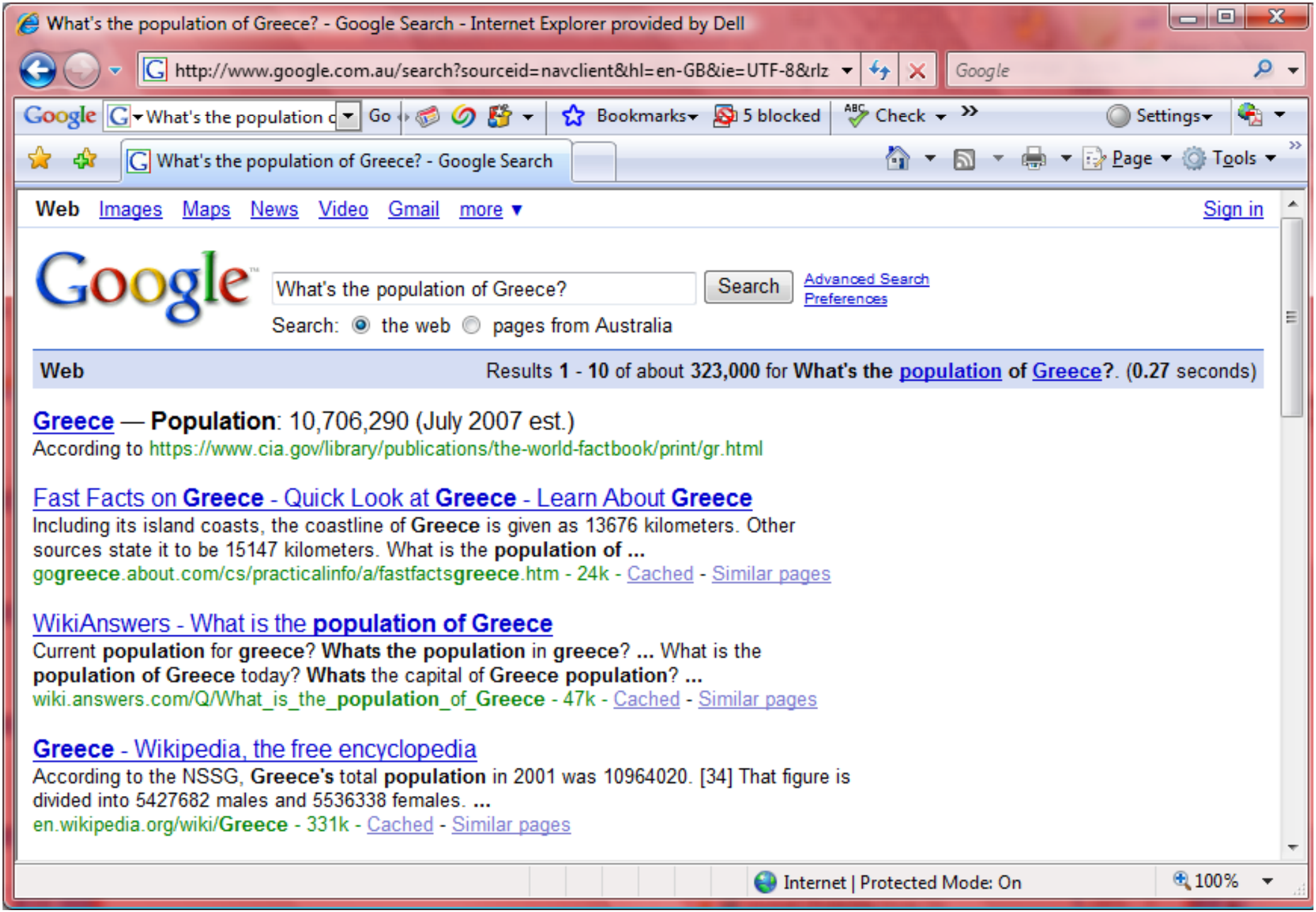
Question-Answering Systems: Current State of the Art

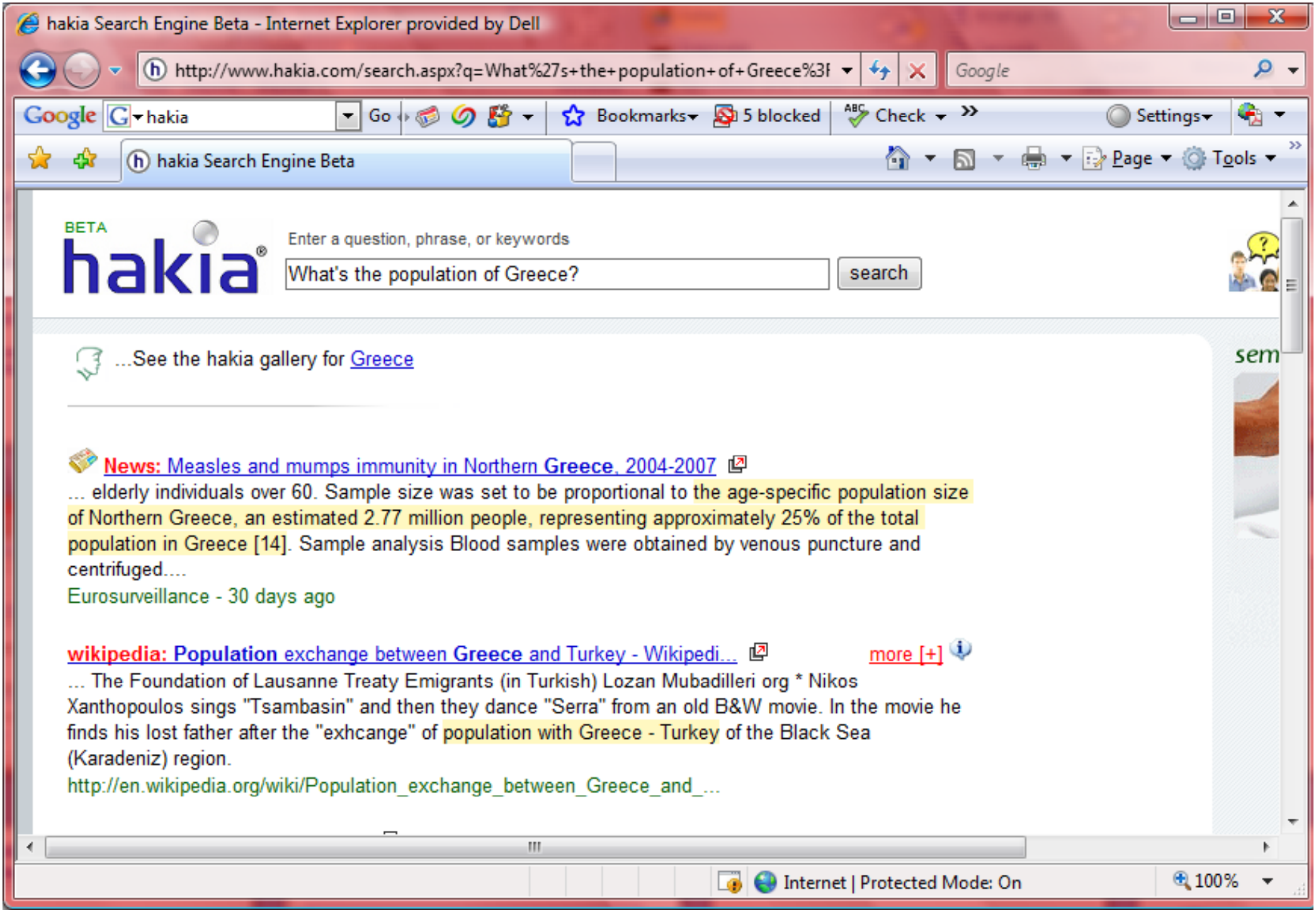
- Limited question analysis to determine query type and central queried concept
- IR techniques used to return appropriate documents
- Data analysis to support construction of custom answers for common questions
- Current technology claimed capable of reducing call center expenses from (for example) $75 a call to 18c a call

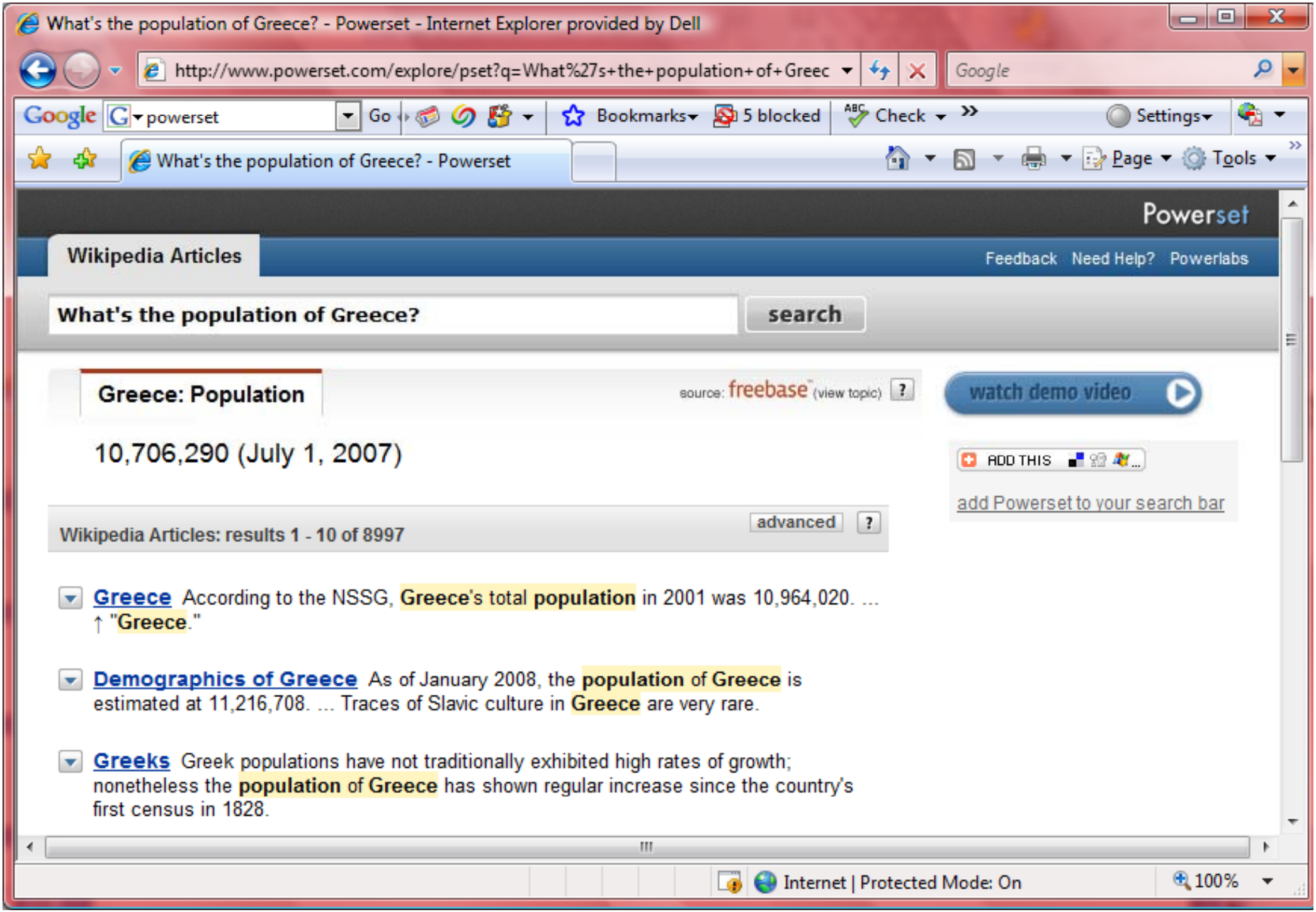
Question-Answering Systems: Fielded Applications

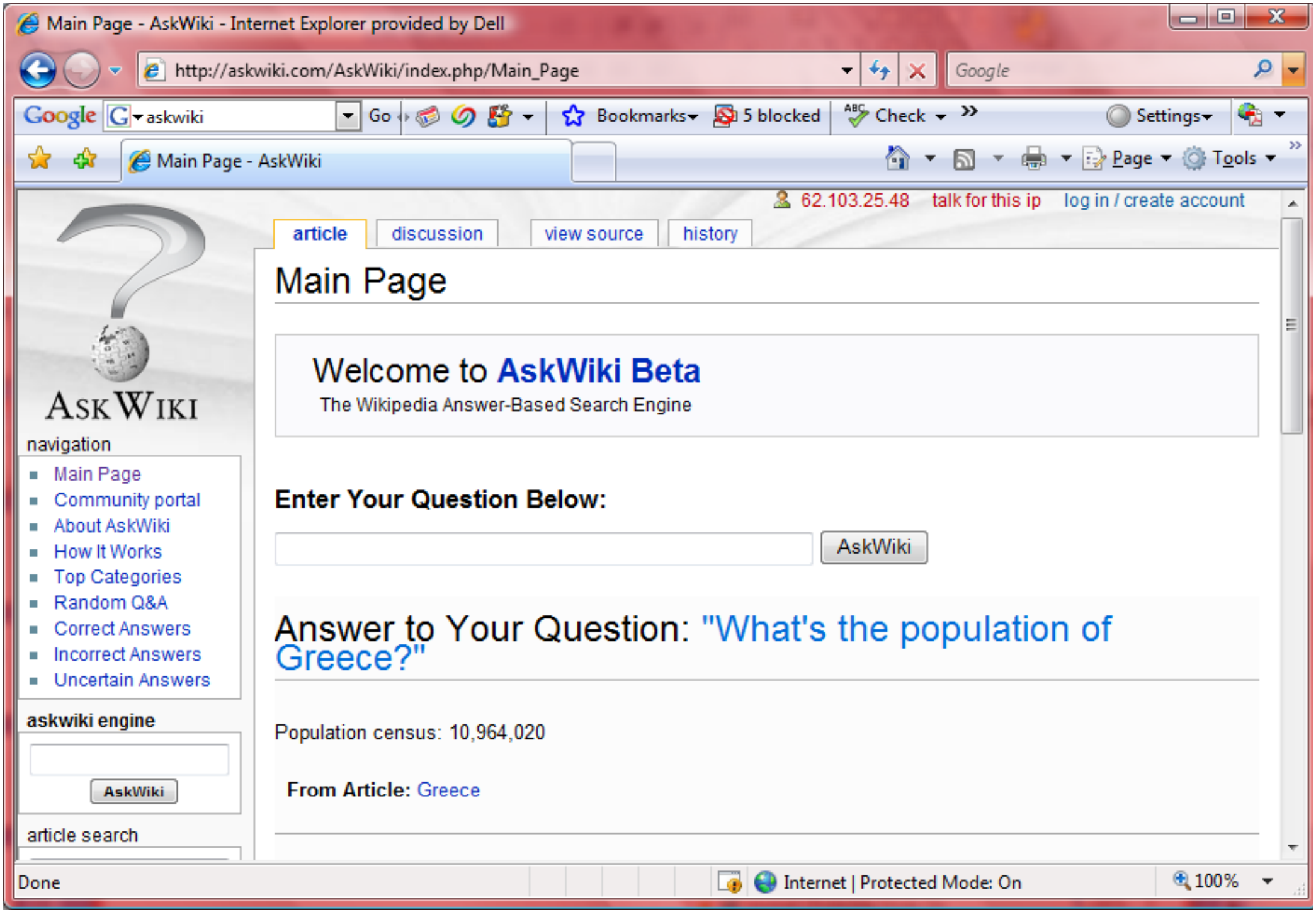
- First generation:
  - Ask Jeeves (www.askjeeves.com)
  - iPhrase Technologies (www.iphrase.com)
  - Soliloquy (www.soliloquy.com)
- Second generation:
  - PowerSet (www.powerset.com)
  - Hakia (www.hakia.com)

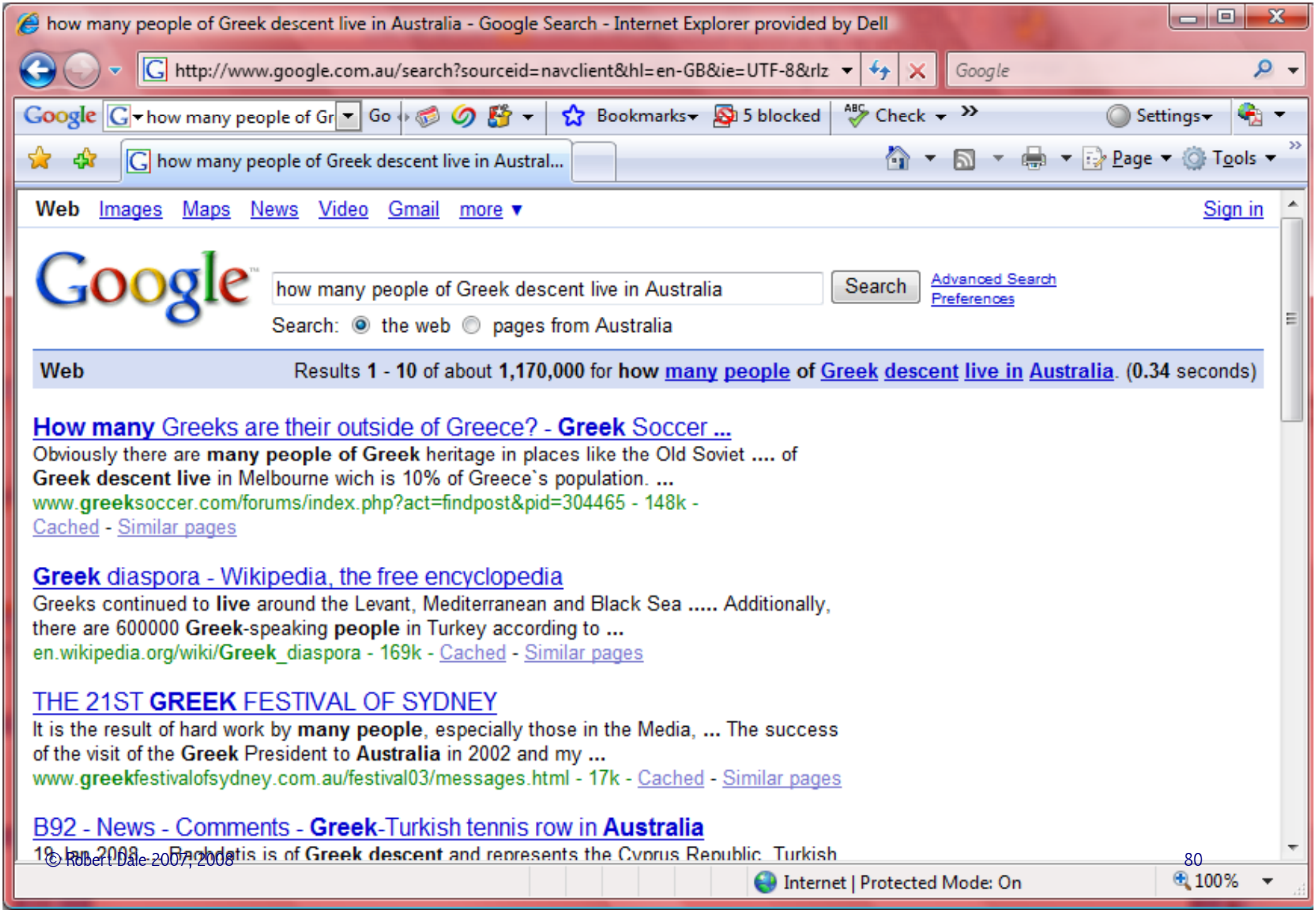

Native Minds - Microsoft Internet Explorer
File Edit View Favorites Tools Help
NativeMinds
Virtual Workforce for the Web
Solutions
Demos
Customers
Partners
Resources
About us
Careers
News
R.O.I.
Services
NativeMinds provides a comprehensive suite of products and services to create automated virtual representatives – vReps – for e-Business customer service, sales, and marketing.
Available 24/7 to answer questions in real-time using natural language, vReps emulate the best in human customer service at a fraction of the cost of traditional support channels such as phone, fax, live chat or email
Hi! My name is Nicole.
May I help you?
Talk to Nicole
Just some of NativeMinds Customers:
ORACLE
Want to see how a vRep works?
QUICK TOUR
DEMOs
In the News
eAssist Partners with NativeMinds to Deploy Self-Service Virtual Representatives for eCRM
Latest from NativeMinds
TA Assoc., Oracle Corp., & CIBC Lead $27.5 Million Venture Financing for NM
July 17 2000
Internet

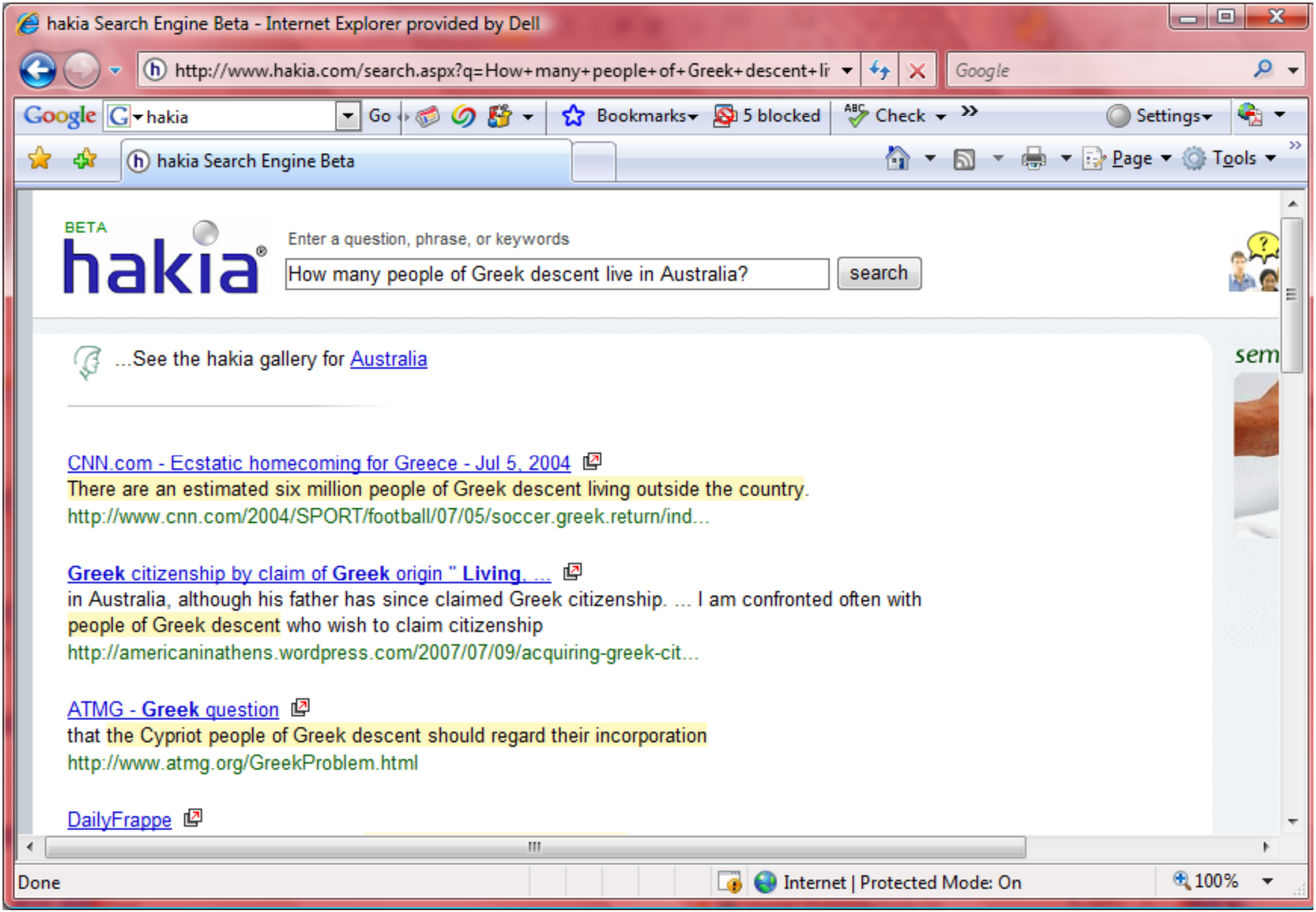

What's the population of Greece? - Google Search - Internet Explorer provided by Dell
http://www.google.com.au/search?sourceid=navclient&hl=en-GB&ie=UTF-8&rlz
Google
What's the population c
Go
Bookmarks
5 blocked
Check
Settings
What's the population of Greece? - Google Search
Page
Tools
Web Images Maps News Video Gmail more ▼
Sign in
Google
What's the population of Greece?
Search
Advanced Search
Preferences
Search: the web pages from Australia
Web
Results 1 - 10 of about 323,000 for What's the population of Greece?. (0.27 seconds)
Greece — Population: 10,706,290 (July 2007 est.)
According to https://www.cia.gov/library/publications/the-world-factbook/print/gr.html
Fast Facts on Greece - Quick Look at Greece - Learn About Greece
Including its island coasts, the coastline of Greece is given as 13676 kilometers. Other sources state it to be 15147 kilometers. What is the population of ...
gogreece.about.com/cs/practicalinfo/a/fastfactsgreece.htm - 24k - Cached - Similar pages
WikiAnswers - What is the population of Greece
Current population for greece? Whats the population in greece? ... What is the population of Greece today? Whats the capital of Greece population? ...
wiki.answers.com/Q/What_is_the_population_of_Greece - 47k - Cached - Similar pages
Greece - Wikipedia, the free encyclopedia
According to the NSSG, Greece's total population in 2001 was 10964020. [34] That figure is divided into 5427682 males and 5536338 females. ...
en.wikipedia.org/wiki/Greece - 331k - Cached - Similar pages
Internet | Protected Mode: On
100%

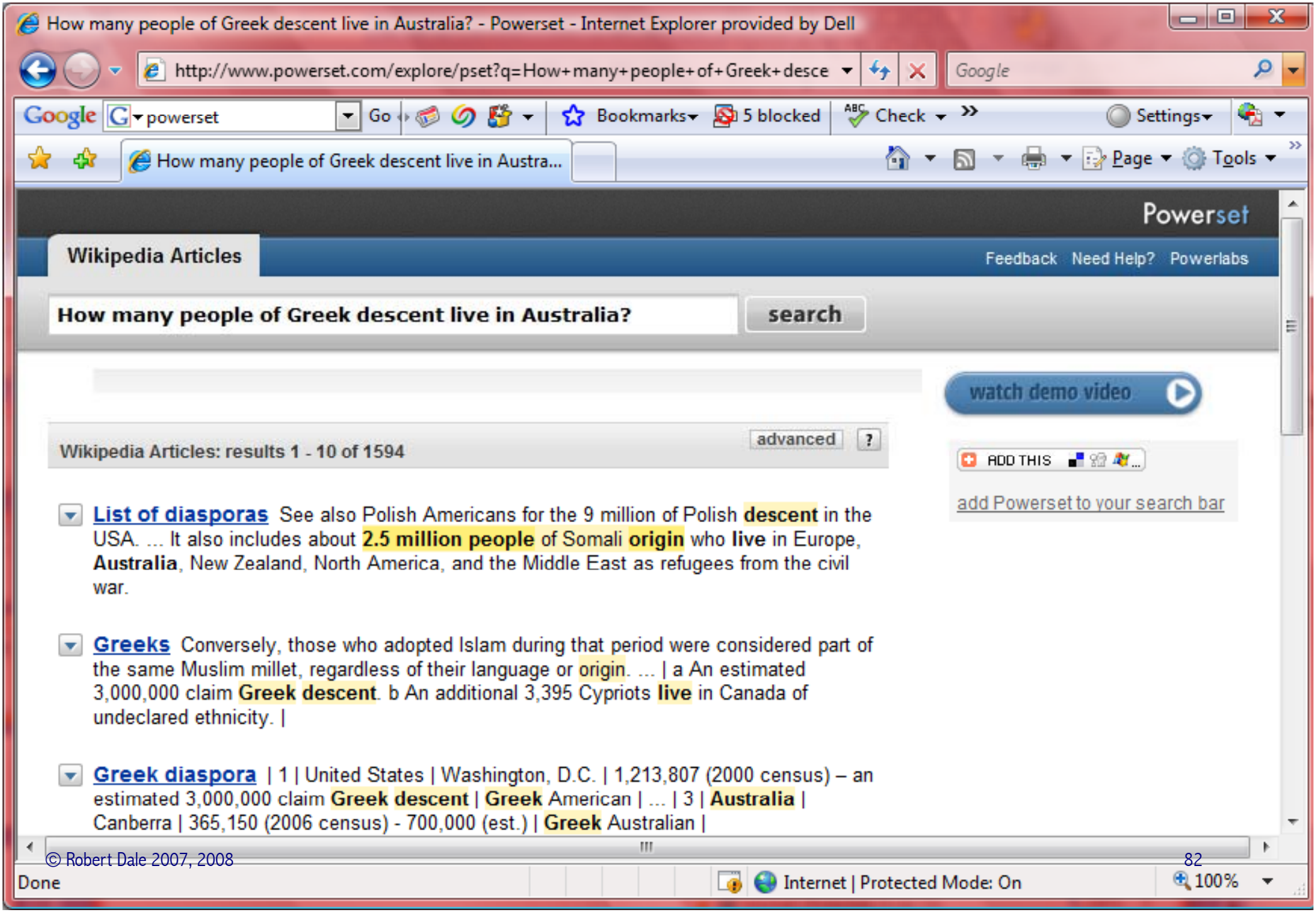

hakia Search Engine Beta - Internet Explorer provided by Dell
http://www.hakia.com/search.aspx?q=What%27s+the+population+of+Greece%3F
Google
hakia
Go
Bookmarks
5 blocked
Check
Settings
hakia Search Engine Beta
Page
Tools
BETA
hakia
Enter a question, phrase, or keywords
What's the population of Greece?
search
...See the hakia gallery for Greece
News: Measles and mumps immunity in Northern Greece, 2004-2007
... elderly individuals over 60. Sample size was set to be proportional to the age-specific population size of Northern Greece, an estimated 2.77 million people, representing approximately 25% of the total population in Greece [14]. Sample analysis Blood samples were obtained by venous puncture and centrifuged....
Eurosurveillance - 30 days ago
wikipedia: Population exchange between Greece and Turkey - Wikipedi...
more [+]
... The Foundation of Lausanne Treaty Emigrants (in Turkish) Lozan Mubadilleri org * Nikos Xanthopoulos sings "Tsambasin" and then they dance "Serra" from an old B&W movie. In the movie he finds his lost father after the "exhcange" of population with Greece - Turkey of the Black Sea (Karadeniz) region.
http://en.wikipedia.org/wiki/Population_exchange_between_Greece_and_...
Internet | Protected Mode: On
100%

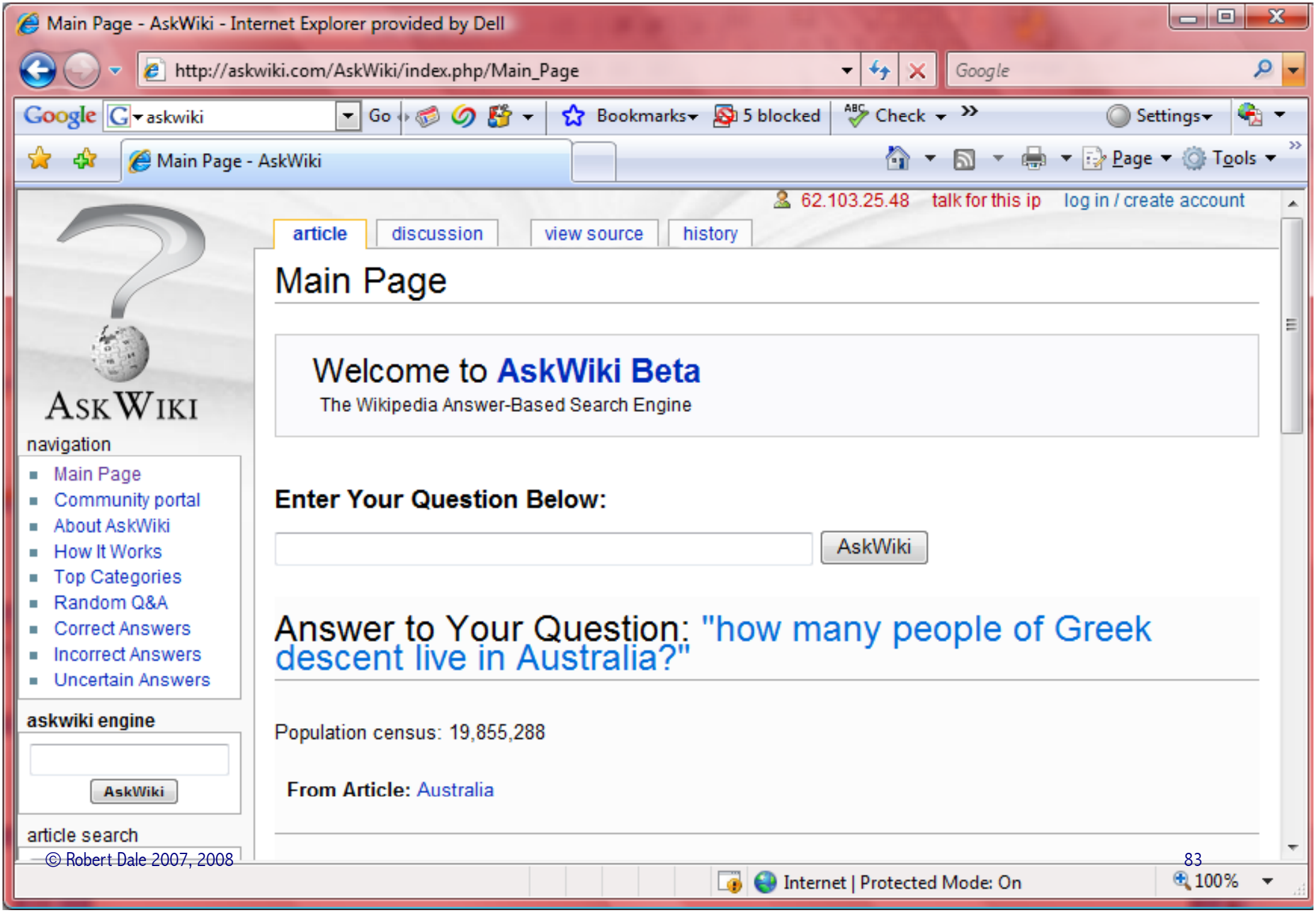

What's the population of Greece? - Powerset - Internet Explorer provided by Dell
http://www.powerset.com/explore/pset?q=What%27s+the+population+of+Greec
Google
powerset
Go
Bookmarks
5 blocked
Check
Settings
What's the population of Greece? - Powerset
Page
Tools
Powerset
Wikipedia Articles
Feedback
Need Help?
Powerlabs
What's the population of Greece?
search
Greece: Population
source: freebase (view topic)
watch demo video
10,706,290 (July 1, 2007)
ADD THIS
add Powerset to your search bar
Wikipedia Articles: results 1 - 10 of 8997
advanced
Greece According to the NSSG, Greece's total population in 2001 was 10,964,020. ... ↑ "Greece."
Demographics of Greece As of January 2008, the population of Greece is estimated at 11,216,708. ... Traces of Slavic culture in Greece are very rare.
Greeks Greek populations have not traditionally exhibited high rates of growth; nonetheless the population of Greece has shown regular increase since the country's first census in 1828.
Internet | Protected Mode: On
100%


Main Page - AskWiki - Internet Explorer provided by Dell
http://askwiki.com/AskWiki/index.php/Main_Page
Google
askwiki
Go
Bookmarks
5 blocked
Check
Settings
Main Page - AskWiki
Page
Tools
62.103.25.48
talk for this ip
log in / create account
article
discussion
view source
history
ASKWIKI
navigation
Main Page
Community portal
About AskWiki
How It Works
Top Categories
Random Q&A
Correct Answers
Incorrect Answers
Uncertain Answers
askwiki engine
AskWiki
article search
Main Page
Welcome to AskWiki Beta
The Wikipedia Answer-Based Search Engine
Enter Your Question Below:
AskWiki
Answer to Your Question: "What's the population of Greece?"
Population census: 10,964,020
From Article: Greece
Done
Internet | Protected Mode: On
100%


how many people of Greek descent live in Australia - Google Search - Internet Explorer provided by Dell
http://www.google.com.au/search?sourceid=navclient&hl=en-GB&ie=UTF-8&rlz
Google
how many people of Gr
Go
Bookmarks
5 blocked
Check
Settings
how many people of Greek descent live in Austral...
Page
Tools
Web Images Maps News Video Gmail more ▼
Sign in
Google
how many people of Greek descent live in Australia
Search
Advanced Search
Preferences
Search: the web pages from Australia
Web
Results 1 - 10 of about 1,170,000 for how many people of Greek descent live in Australia. (0.34 seconds)
How many Greeks are their outside of Greece? - Greek Soccer ...
Obviously there are many people of Greek heritage in places like the Old Soviet .... of Greek descent live in Melbourne wich is 10% of Greece`s population. ...
www.greeksoccer.com/forums/index.php?act=findpost&pid=304465 - 148k -
Cached - Similar pages
Greek diaspora - Wikipedia, the free encyclopedia
Greeks continued to live around the Levant, Mediterranean and Black Sea ..... Additionally, there are 600000 Greek-speaking people in Turkey according to ...
en.wikipedia.org/wiki/Greek_diaspora - 169k - Cached - Similar pages
THE 21ST GREEK FESTIVAL OF SYDNEY
It is the result of hard work by many people, especially those in the Media, ... The success of the visit of the Greek President to Australia in 2002 and my ...
www.greekfestivalofsydney.com.au/festival03/messages.html - 17k - Cached - Similar pages
B92 - News - Comments - Greek-Turkish tennis row in Australia
19 Jan 2008 ... Baghdatis is of Greek descent and represents the Cyprus Republic. Turkish
Internet | Protected Mode: On
100%


hakia Search Engine Beta - Internet Explorer provided by Dell
http://www.hakia.com/search.aspx?q=How+many+people+of+Greek+descent+li
Google
hakia
Go
Bookmarks
5 blocked
Check
Settings
hakia Search Engine Beta
Page
Tools
BETA
hakia
Enter a question, phrase, or keywords
How many people of Greek descent live in Australia?
search
...See the hakia gallery for Australia
CNN.com - Ecstatic homecoming for Greece - Jul 5, 2004
There are an estimated six million people of Greek descent living outside the country.
http://www.cnn.com/2004/SPORT/football/07/05/soccer.greek.return/ind...
Greek citizenship by claim of Greek origin " Living, ...
in Australia, although his father has since claimed Greek citizenship. ... I am confronted often with people of Greek descent who wish to claim citizenship
http://americaninathens.wordpress.com/2007/07/09/acquiring-greek-cit...
ATMG - Greek question
that the Cypriot people of Greek descent should regard their incorporation
http://www.atmg.org/GreekProblem.html
DailyFrappe
Done
Internet | Protected Mode: On
100%





Principal Components in a Language Technology Application

- Language input
  - recognizing the words
- Language processing
  - reasoning about the words to get at their meaning
- Language output
  - rendering meaning as words

Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation

Text-to-Speech: Key Focus and Applications

- Key Focus of the Technology
  - Production of natural sounding speech from a textual input
- Applications
  - Spoken rendering of email via desktop and telephone
  - Document proofreading
  - Voice portals

Text-to-Speech: Issues and State of the Art

- TTS in a vacuum requires reverse engineering of linguistic information
  - Appropriate use of intonation and phrasing
  - Handling homophones
- High quality diphone concatenation is readily available:
  - Short digital-audio segments are concatenated, and intersegment smoothing performed to produce a continuous sound
  - Very appropriate where audio prerecording not usable

Text-to-Speech: Fielded Applications

- Nuance RealSpeak (www.nuance.com/realspeak/)
- Cepstral (www.cepstral.com)
- AT&T Natural Voices (www.research.att.com/~ttsweb/tts/index.php)
- For comparison:
  - Diphone synthesis
  - Unit selection

Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation

Tailored Document Generation: Key Focus and Applications

- Key Focus of the Technology
  - Production of individually-tailored documents based on parameter values
- Applications
  - Individual, personalised advice-giving
  - Customised personnel and policy manuals
  - Web-delivered dynamic documents

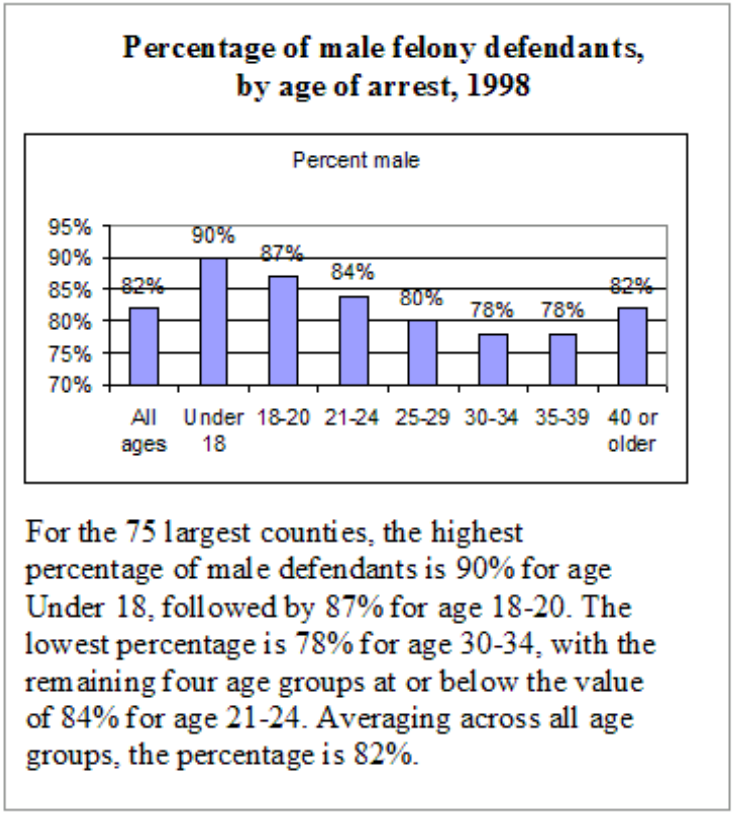
Tailored Document Generation: Issues and State of the Art

- Mail-merge is the bottom-end of this technology
- Tailored composition of document components and associated template filling can produce wide variations in output
- Going beyond mail-merge requires underlying knowledge source rich enough to drive sophisticated linguistic abilities
- Applications with complex underlying models such as project management software or CAD software can provide appropriate input

Tailored Document Generation: Fielded Applications

- KnowledgePoint (www.knowledgepoint.com)
  - Tailored job descriptions and personnel policies
  - Automated performance review systems
- CoGenTex (www.cogentex.com)
  - Automatic generation of project status reports
  - Automatic generation of summaries of charts and tables

CoGenTex: Chart Explainer



For the 75 largest counties, the highest percentage of male defendants is 90% for age Under 18, followed by 87% for age 18-20. The lowest percentage is 78% for age 30-34, with the remaining four age groups at or below the value of 84% for age 21-24. Averaging across all age groups, the percentage is 82%.

Summary

Language Input	Language Processing	Language Output
Speech Recognition	Spoken Language Dialog Systems	Text-to-Speech
Optical Character Recognition	Search and Information Retrieval	Tailored Document Generation
Handwriting Recognition	Writing Assistance	
	Machine Translation	
	Text Summarization	
	Question-answering Systems	

Summary So Far

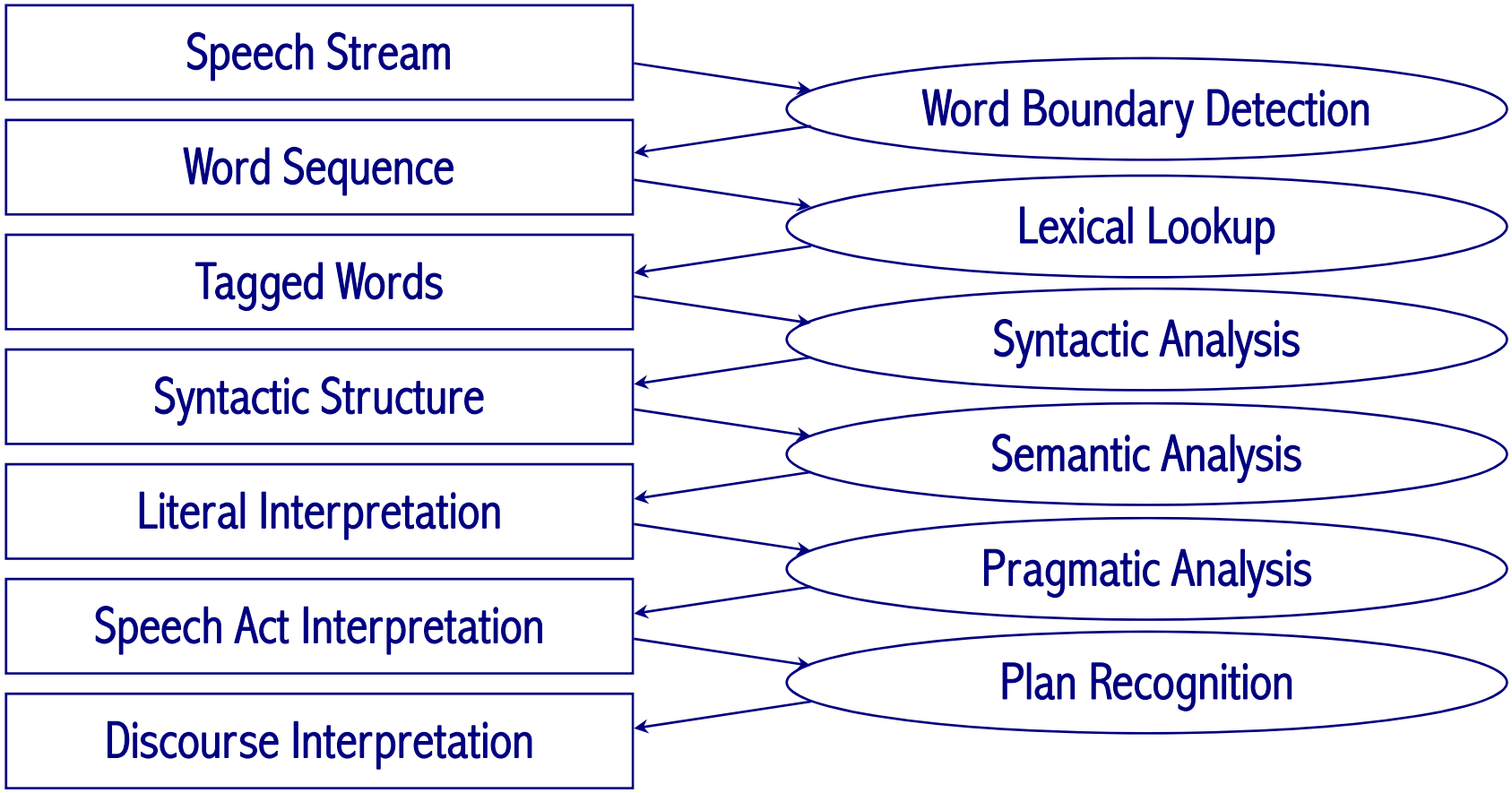
- Input technologies can achieve in excess of 90% accuracy
- Broad coverage applications have to rely on limited linguistic knowledge
- Targetted applications can use more sophisticated linguistic knowledge
- Output technologies are not yet a major focus

Part 2: Techniques

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
- Conclusions and Further Information

Stages in Processing Language



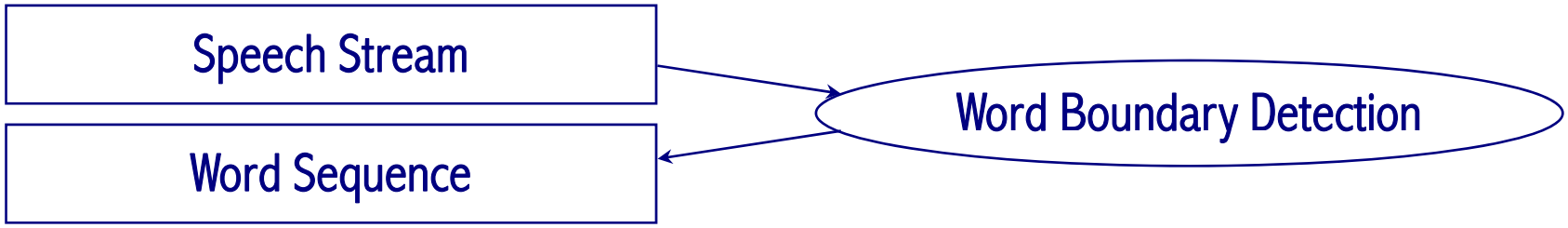
Word Boundary Detection

- recognise speech
- wreck a nice peach

Word Boundary Detection

- A speech recognition system needs to recognise the phonemes that were spoken and then assemble these into valid sequences of words
- Different people pronounce phonemes in different ways: an acoustic model captures a representation of the possible renderings of phonemes that can be matched against
- A language model indicates what sequences of words are possible

Stages in Processing Language



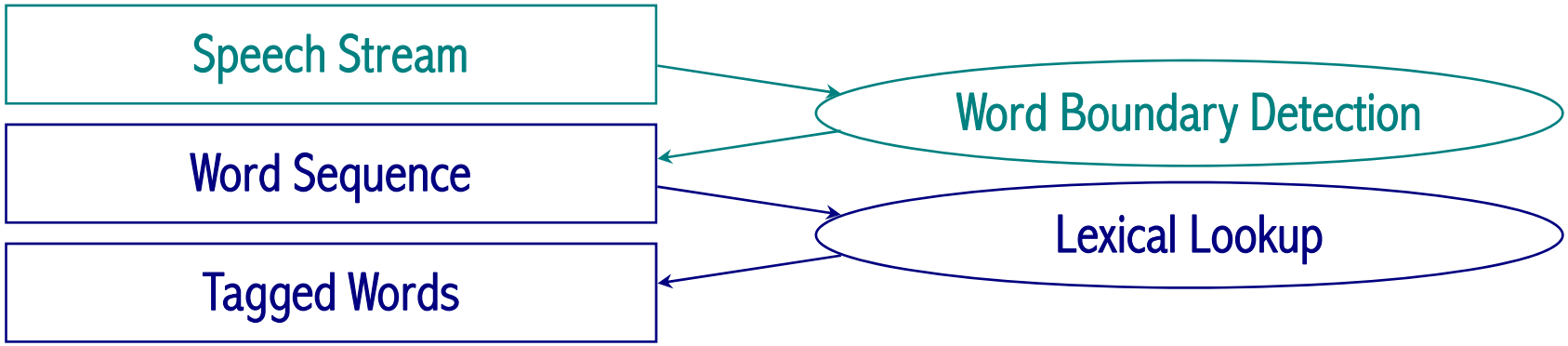
Lexical Ambiguity

- The astronomer saw the star.
- The astronomer married the star.
- King Kong sat on the bank.

Lexical Ambiguity

- Early methods were rule-based and relied on at least a partial understanding of the context
- Selectional restrictions in the lexicon:
  - marry[agent=animate, object=animate]
  - $star_1$[+animate] % famous or celebrated-person
  - $star_2$[–animate] % celestial object
- Modern techniques rely on statistical evidence derived from large bodies of text

Stages in Processing Language

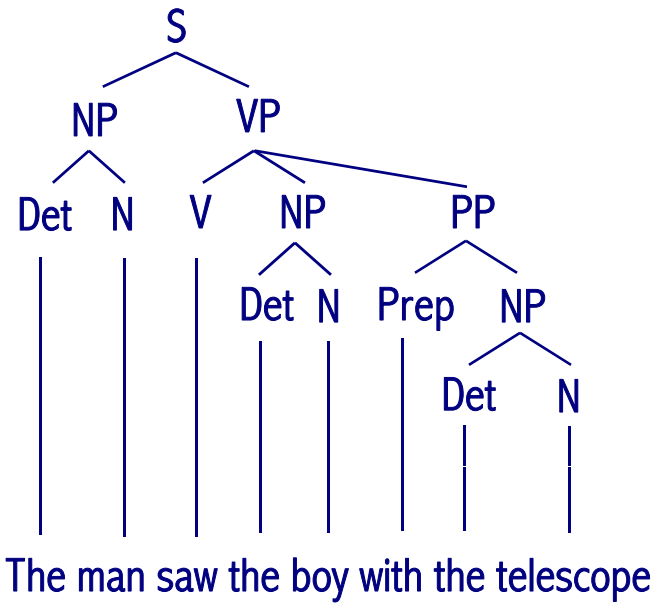


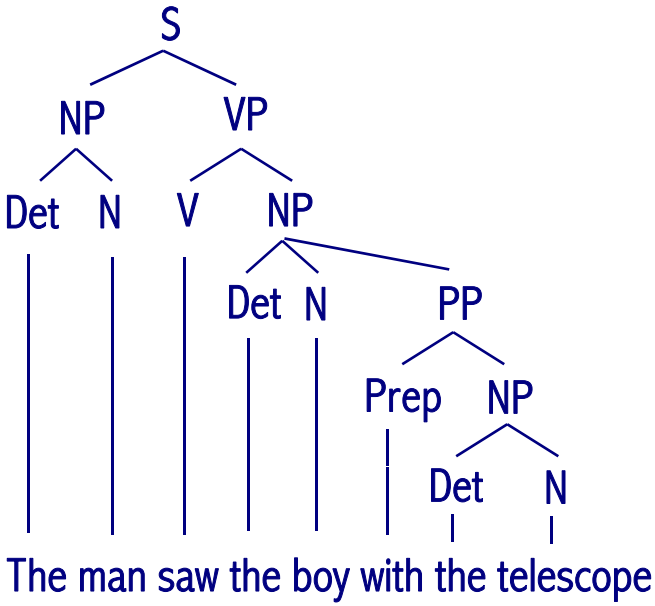
Structural Ambiguity

- The astronomer saw the star with a telescope.
- The astronomer married the star with a history.
- Visiting uncles can be a nuisance.
- I forgot how good beer tastes.

Structural Ambiguity

- The man saw the boy with the telescope





Structural Ambiguity

- A grammar inventorises the possible syntactic structures in a language by means of a fine set of rules
- These rules dictate how symbols in the language can be combined to create well-formed sentences

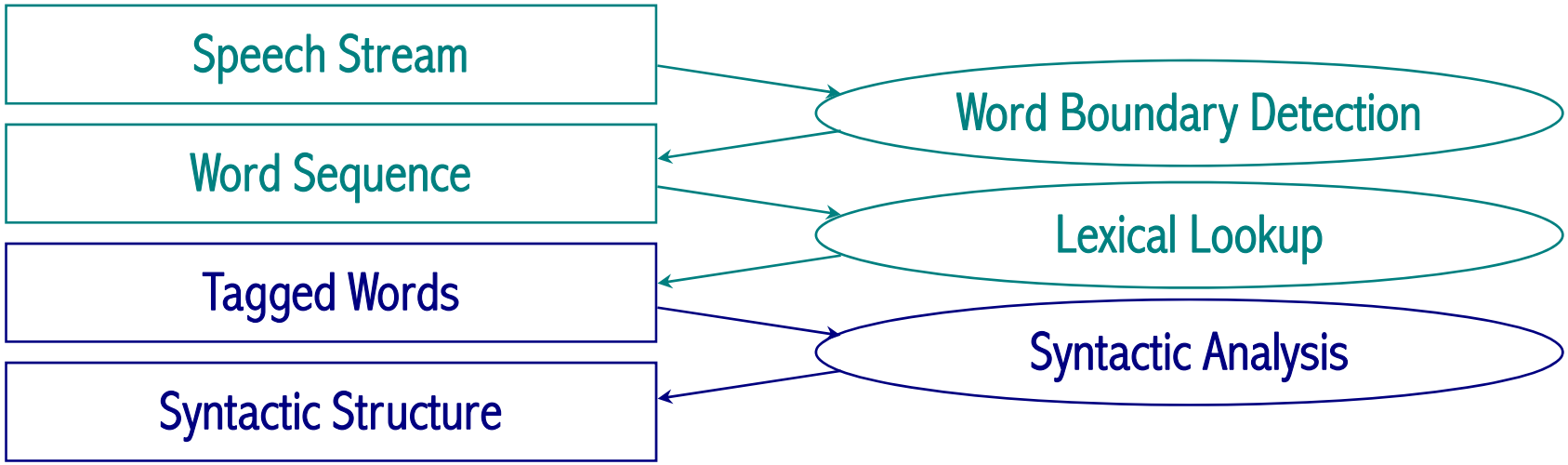
  $S \rightarrow NP\ VP$

  $NP \rightarrow Det\ N$

  $VP \rightarrow V\ NP$

- A parser uses a set of grammar rules to attribute a syntactic structure to a well-formed string

Stages in Processing Language



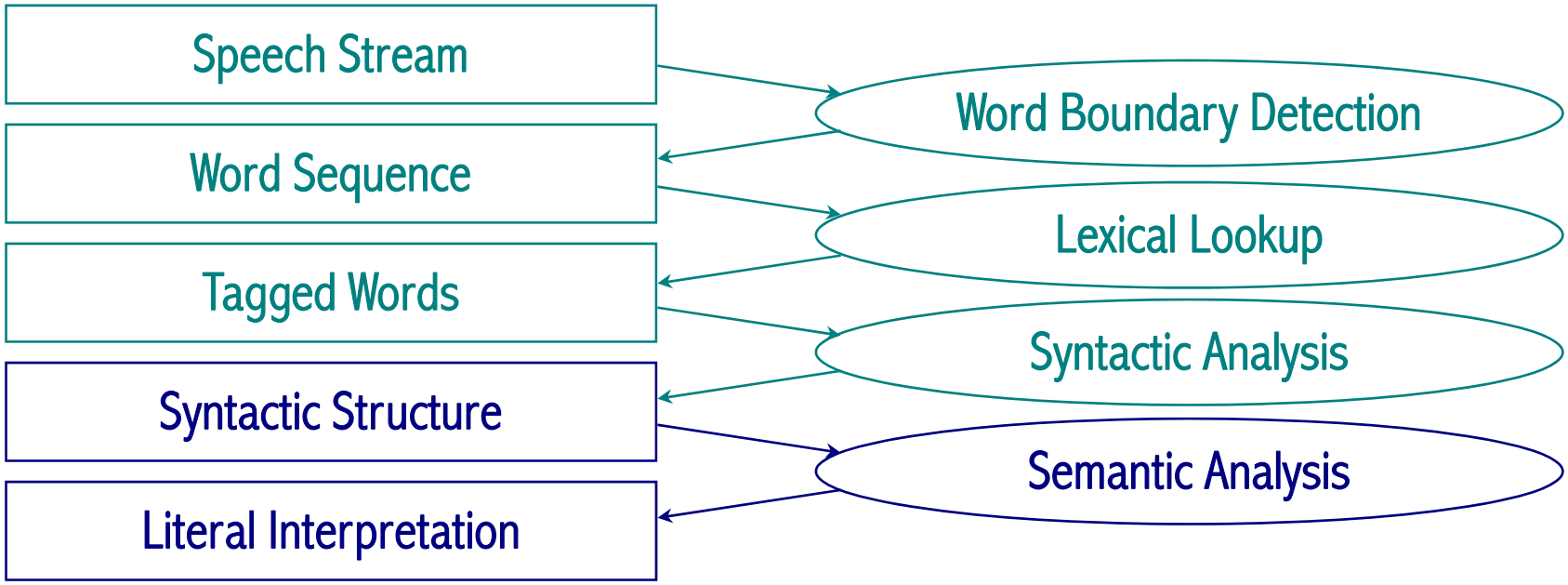
Anaphora Resolution

- The councillors refused the women a permit because they feared revolution.
- The councillors refused the women a permit because they advocated revolution.

Anaphora Resolution

- Anaphora resolution is just one of a range of problems in semantic interpretation
- Anaphora resolution involves all kinds of linguistic knowledge: intonational, syntactic, semantic and pragmatic:
  - Maisy swore at Sabine then she insulted her.
  - Jim hurt him.
  - Andy put the cake on the table and ate it.
  - Sue went to Mary's house and she cooked her dinner.

Stages in Processing Language



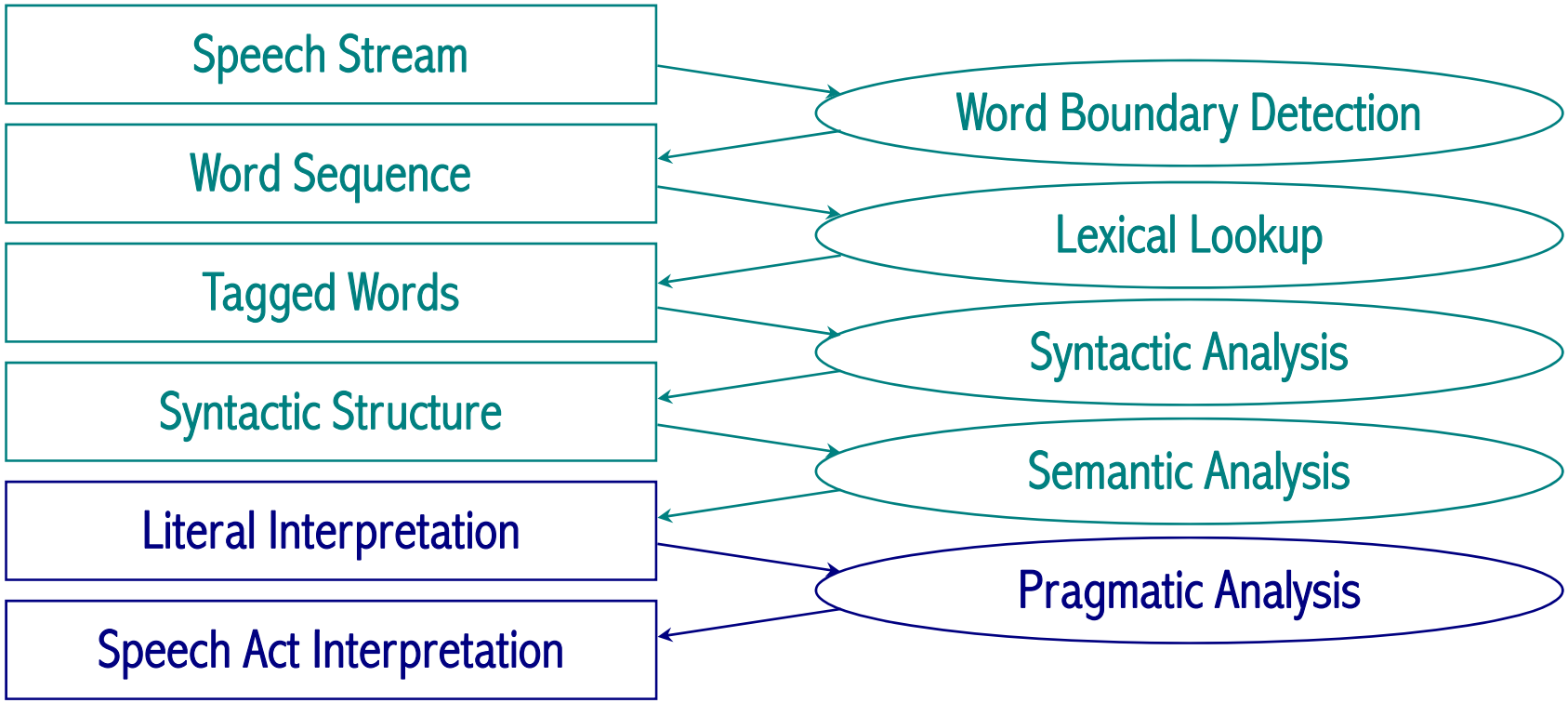
Non-literal Meaning

- Can you pass the salt?
- You're standing on my foot.
- His handwriting is very good.

Non-literal Meaning

- We always understand language in a context
- Our rich store of world knowledge allows us to draw the appropriate inferences to construct an appropriate interpretation
- Access to a similar store of world knowledge is a significant problem for computers
- As a result, successful applications of NLP lie in areas where we can closely constrain the context and therefore the range of possible interpretations

Stages in Processing Language

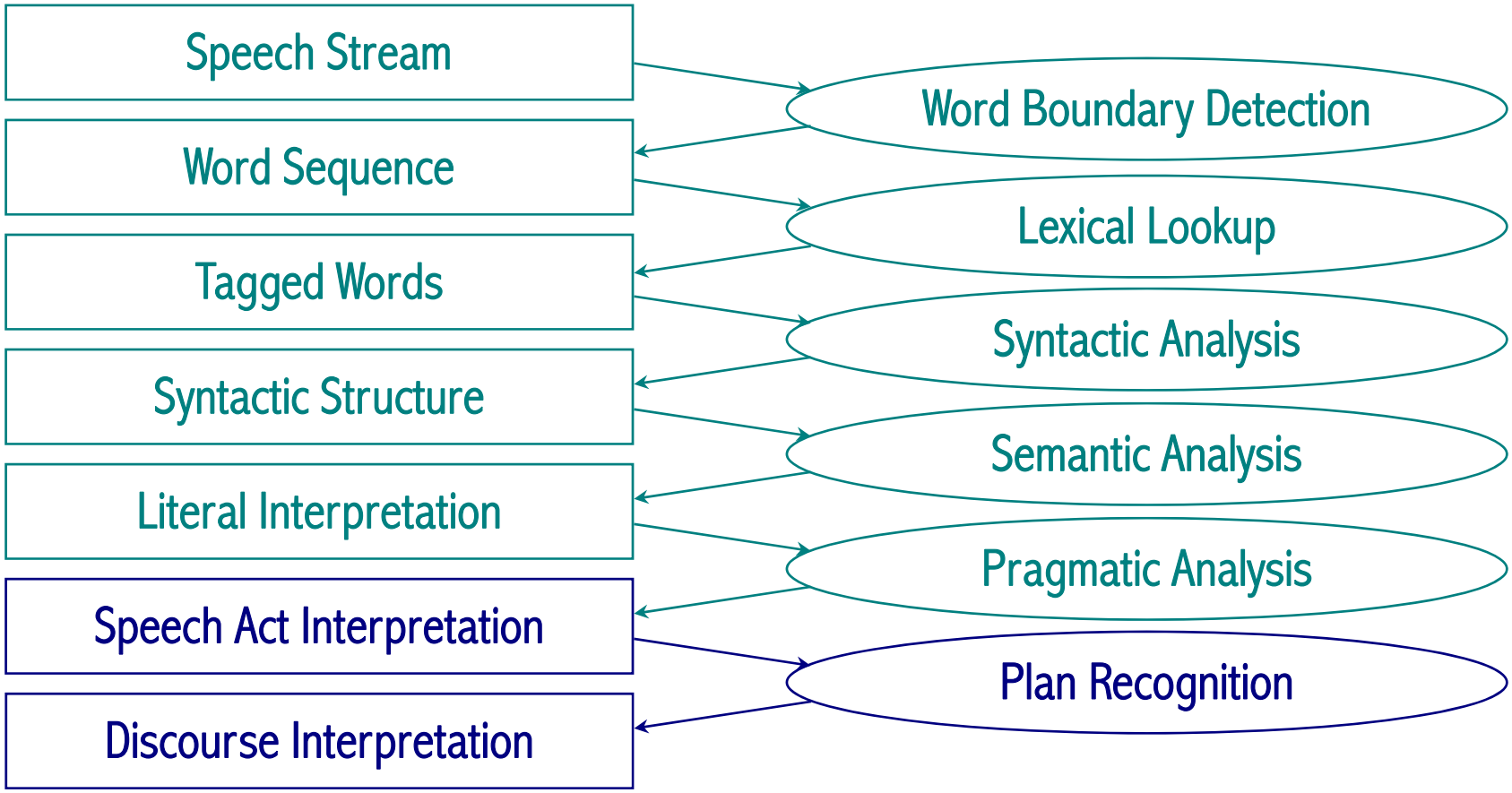


Plan Recognition

Plan inference and co-operative response:

User:	*Which students got an F in Comp248 in 1993?*
System:	None.
User:	*Did anyone fail Comp248 in 1993?*
System:	No.
User:	*How many people passed Comp248 in 1993?*
System:	Zero.
User:	*Was Comp248 given in 1993?*
System:	No.

Stages in Processing Language



Plan Recognition

- When we take part in dialog, we are constantly making predictions as to what the other party in the dialog wants
- Research systems use complex inferences over assumed user beliefs and intentions
- Truly intelligent systems need to do the same thing
- Meaning results from the text and the context in combination

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Getting Language into the Machine

- **Speech Stream:** recognize phonemes, segment into words, represent as a stream or lattice of space-separated word tokens
- **Handwriting Recognition:** recognise characters in cursive script, represent as space-separated word tokens
- **Optical Character Recognition:** recognise characters within page layout, combine into space-separated word tokens
- **Existing Electronically Encoded Documents:** strip out formatting commands and control characters, represent as space-separated word tokens

Getting Language into the Machine

- Tokenisation:
  - the process of breaking up a sequence of characters in a text by locating the word boundaries
  - the words thus identified are tokens
  - in languages where no word boundaries are explicitly marked in the writing system, this is also known as word segmentation

Getting Language into the Machine

- Sentence Segmentation
  - the process of identifying sentence boundaries
  - involves sentence boundary detection, disambiguation or recognition

Tokenisation and Sentence Segmentation

- The two tasks are not independent:
  - Maria finished her Ph.D. yesterday.
  - Yesterday Maria finished her Ph.D.
- Real sentence boundary recognition is hard!
  - Two high-ranking positions were filled Friday by Penn St. University President Graham Spencer.
  - Two high-ranking positions were filled Friday by Penn St. University President Graham Spencer announced the appointments.

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Word Lists

- The minimal linguistic resource required for many applications: a list of the words in the language
  - Generally required for spell checking and correction
  - Can reduce error rates in OCR and handwriting recognition
- Spell checking can also be carried out using lists of valid character bigrams or trigrams—but this isn't enough for correction

Word Lists

- Existing IR and Text Summarisation systems can perform without word lists:
  - In simple IR, words are just strings of characters
  - In simple Text Summarisation, sentences are just sequences of words, which are strings of characters
- Benefit: absolutely broad coverage
- Cost: zero leverage of linguistic information

Word Lists

- A typical desk dictionary contains around 50000–150000 entries
- In 44 million words of Associated Press newswire text collected over 10 months, there were 300000 different tokens
- How do you build a lexicon big enough to deal with real language?
- One possibility: make use of machine readable dictionaries
- A popular MRD: Longman's Dictionary of Contemporary English

Word Lists

- How many words do you need? It has been suggested that by age 17 we know 80000 words.
- But: it has been estimated that 8000 base forms of words (morphemes) is sufficient to handle 95% of texts
- Typically, 15 most frequent words account for 25% of tokens
- 100 most frequent words account for 60% of tokens

Word Frequencies

Rank	Spoken English	Written English	French	German	
1	the	the	de	der	
2	and	of	le	die	
3	I	to	la	und	
4	to	in	et	in	
5	of	and	les	des	
6	a	a	des	den	
7	you	for	est	zu	
8	that	was	un	das	
9	in	is	ure	von	
10	it	that	du	fur	

Dictionaries

- A dictionary (or lexicon) is a collection of words with associated information:
  - A mapping to phonetic transcriptions is required for speech recognition
  - A mapping to parts of speech is required for almost all language technology applications that do anything with the words once recognised

Dictionaries: Phonetic

- The Roman alphabet has 26 characters, but English has around 44 distinct phonemes
- Phonetic transcription traditionally notated using IPA, the International Phonetic Alphabet, but more recent encodings are computer-readable

ðiɪntəˈnæʃənəlfəˈnɛtɪkəsoʊsiˈeɪʃn

Dictionaries: Part of Speech

- Every word has a Syntactic Category or Part of Speech
- Parts of speech are important because they constrain how sentences can be put together
- Two broad types: Open Class words vs Closed Class words
- This information is needed for syntactic analysis
- Problem: dealing with unknown words

Dictionaries: Part of Speech

- Nouns
  - projector, money, infidelity, amazement, antidisestablishmentarianism …
- Verbs
  - run, fly, walk, procrastinate, believe …
- Adjectives
  - crazy, green, hungry, unbelievable, amazed, smart …
- Adverbs
  - slowly, hungrily, unbelievably …

Dictionaries: Part of Speech

- Determiners
  - a, the, this, that, these, those …
- Conjunctions
  - and, but, therefore, because …
- Prepositions
  - in, on, under, between, to, from …

Morphology and the Dictionary

- Listing information on every word in the language separately fails to observe that there are systematic relationships between words
- We can save space by recognising the morphological structure of words, and constructing them from their component parts by rule
- Morphological processing can help in providing Part of Speech information for unknown words

Inflectional Morphology

- Root Form + Affix; affix can be a Prefix, Infix or Suffix
- Part of speech remains constant; same basic meaning
- Examples:
  - deliver + s = delivers [third person singular present tense]
  - deliver + ing = delivering [present participle]
  - deliver + ed = delivered [past tense]
- Root form also known as the Base, Stem, or Lemma
- Root forms are Free Morphemes
- Affixes are usually Bound Morphemes

Derivational Morphology

- A word of one category is used to derive a word of another category
- friend [noun] + ly [suffix] = friendly [adjective]
- friendly [adjective] + ness [suffix] = friendliness [noun]

Stemming

- Many IR systems use a linguistically under-motivated but simpler process called stemming to conflate words with a common base
- A popular model: the Porter Stemmer

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Building Syntactic Representations

- A significant proportion of the work in traditional NLP has focused on syntactic analysis
  - sophisticated linguistic formalisms for capturing generalisations
  - efficient parsing techniques for broad coverage syntactic analysis

Applications of Syntactic Analysis

- Rich analysis generally required for
  - Grammar checking
  - Transfer-based and Interlingua-based Machine Translation

Applications of Syntactic Analysis

- Limited syntactic coverage required for:
  - Spoken-language dialog systems
  - Question-answering systems

Applications of Syntactic Analysis

- Shallower techniques based on finite state grammars sufficient for
  - Concept-based information retrieval
  - Information extraction technologies

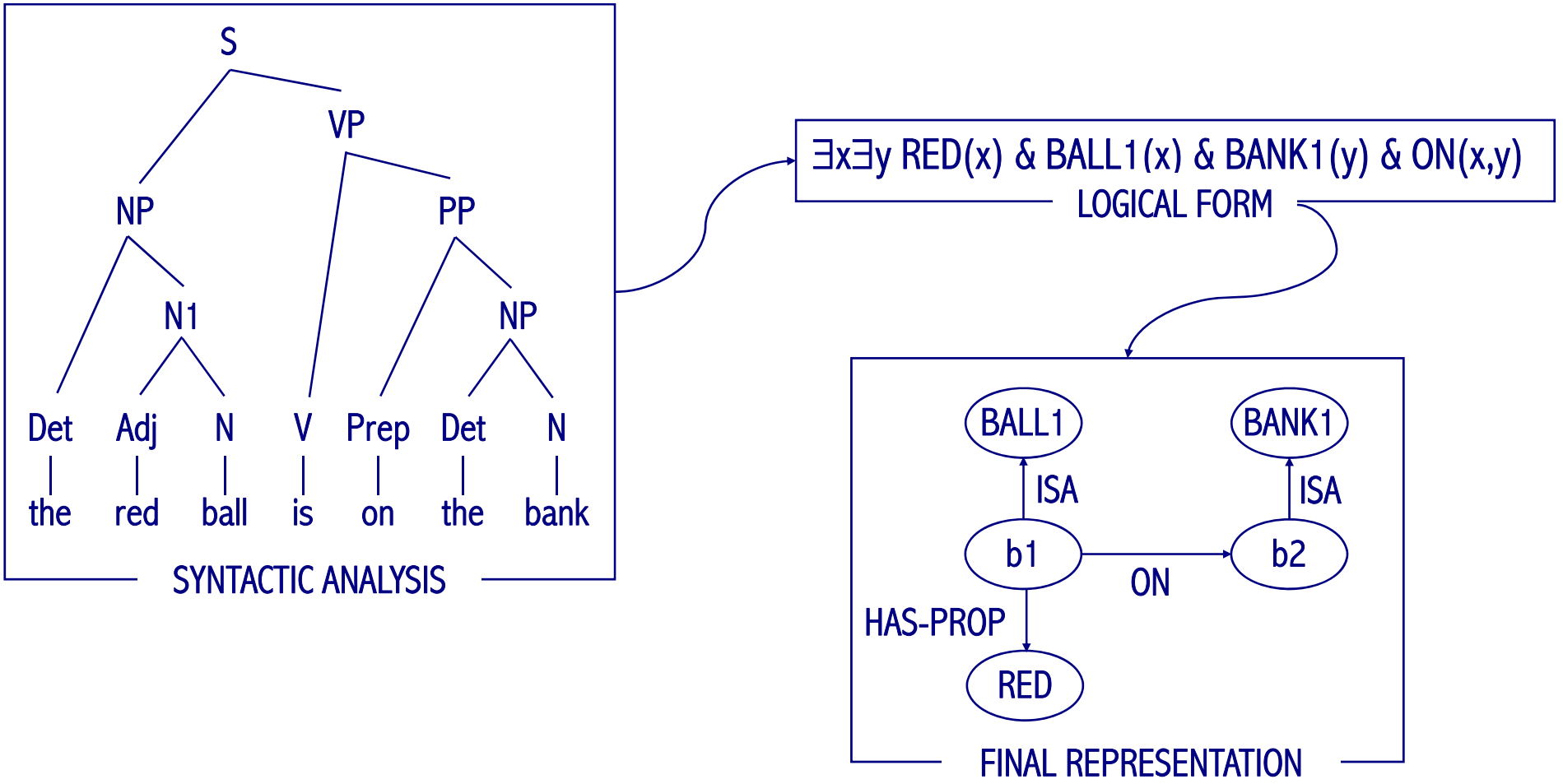
Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Semantics as Logical Form

- Typically expressed using First Order Predicate Calculus:
  - variables
  - predicates
  - logical connectives
  - quantifiers
- Other forms of logic required to express possibility, necessity, temporal phenomena …

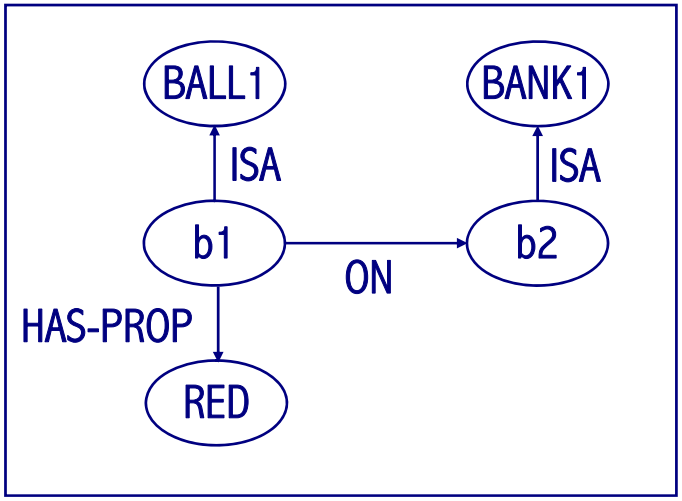
From Syntax to Semantics



From Syntax to Semantics

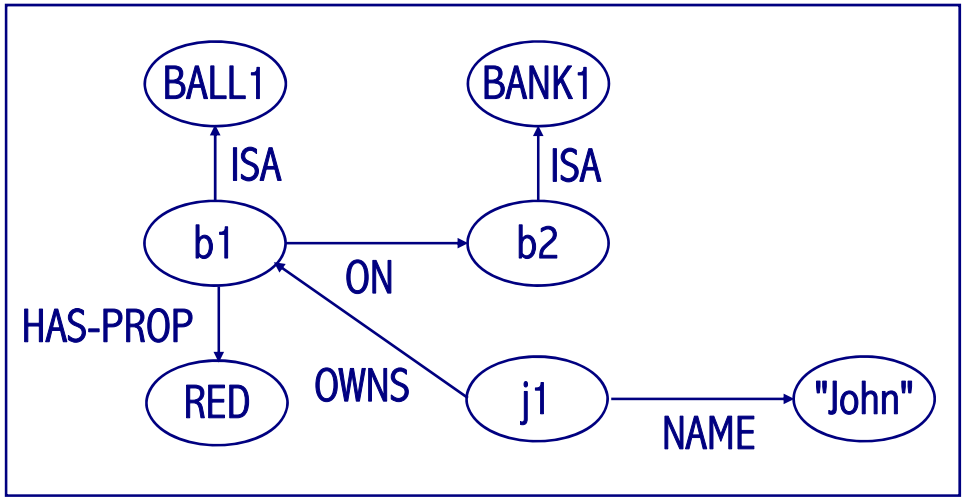
The red ball is on the bank.

$\exists x \exists y$ RED(x) & BALL1(x) & BANK1(y) & ON(x,y)



It belongs to John.

$\exists x \exists y$ NAME(x, John) & OWNS(x,y)



How Do We Get From Syntax to Semantics?

- Meaning is compositional: the meaning of a constituent is derived solely from the meanings of its subconstituents and their means of combination
- An elegant approach: the lambda calculus
- Each lexical entry expresses the meaning of the word as a lambda expression; the rules of the grammar indicate how these expressions are to be combined

Case Frames

- If we ignore quantificational phenomena, most significant aspect of meaning is 'who did what to whom'
- Semantically, each verb carries a set of case roles that specify the semantic relationships corresponding to the different participants in the event described:
  - AGENT
  - PATIENT
  - INSTRUMENT
  - SOURCE
  - DESTINATION
  - ...

Case Roles and Case Frames

We can introduce events as logical variables:

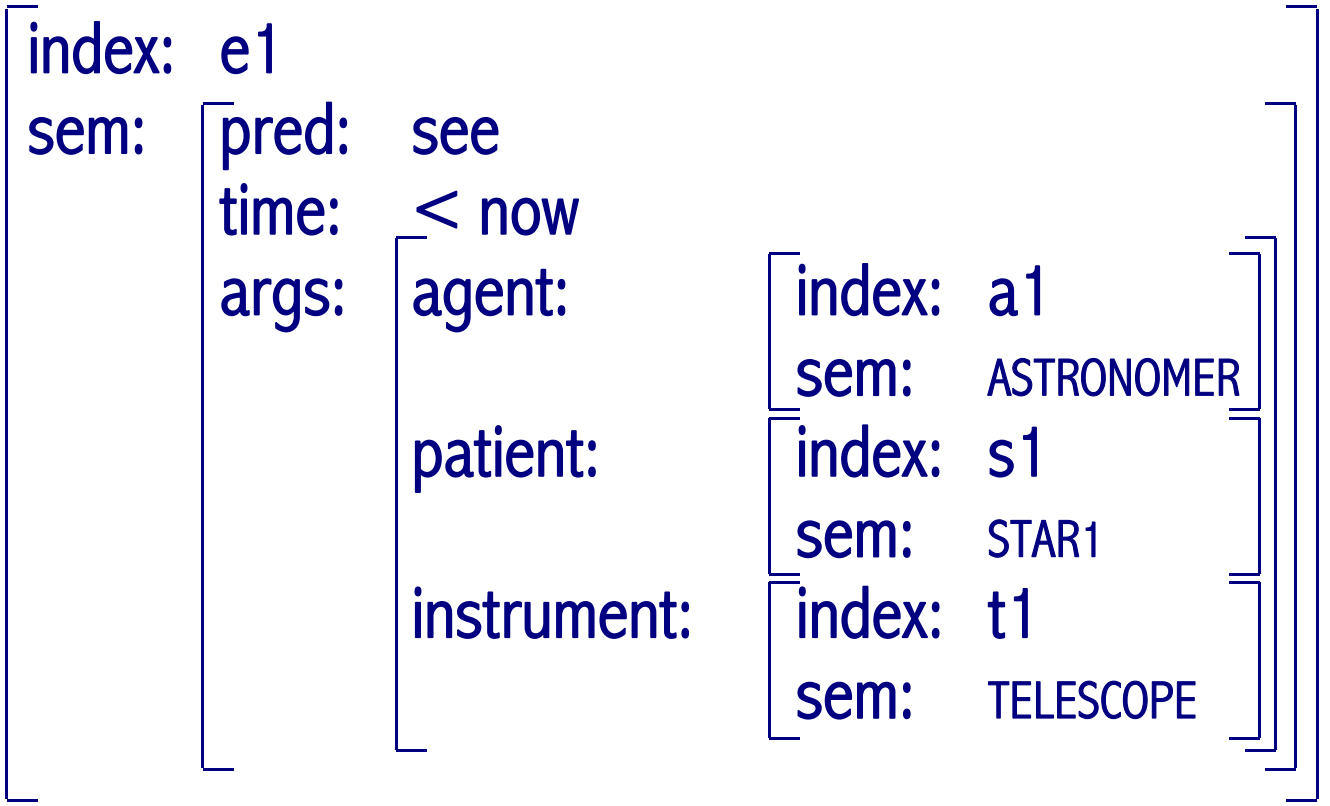
The astronomer saw the star with a telescope

$\exists e \exists x \exists y \exists z$ SEE(e) & PAST(e) & ASTRONOMER(x) & STAR1(y) & TELESCOPE(z) &
AGENT(e,x) & PATIENT(e,y) & INSTRUMENT(e,z)

The astronomer married the star with a birthmark

$\exists e \exists x \exists y \exists z$ MARRY(e) & PAST(e) & ASTRONOMER(x) & STAR2(y) & BIRTHMARK(z) &
AGENT(e,x) & PATIENT(e,y) & POSSESS(y,z)

A Feature Structure Representation



Semantic and Pragmatic Knowledge

- From a theoretical perspective, semantics and pragmatics are distinct
- In practical systems, pragmatic issues are often 'compiled-down' into semantics, or even into the syntax
- For practical applications this is valid because of the limited coverage required

A Grammar Rule in a Dialog System

- Semantics compiled into syntax:

balance-request →
((what is | what's) (my | the) balance [please]) |
((tell me the | check my) balance [please])
<requesttype=balance>

pin-change-request →
((can I | I'd like to) change my (pin [number] | personal identification number | access code) [please]
<requesttype=pin-change>

Interlingua Mappings in Machine Translation

- Representations similar to case frames serve as interlingua: a level of representation that embodies the basic concepts in a language-independent form
- Pragmatics? Some options
  - Pragmatics compiled into semantics
  - Pragmatics as a free lunch
  - Treat special cases separately

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
- Conclusions and Further Information

Technology Map: Spoken Language Dialog Systems

- Limited grammatical coverage: simple syntax, effectively represented by means of semantic grammars
- Rich phonetically-annotated lexica for speech recognition and synthesis
- Hard-wired, implicit pragmatics

Technology Map: Search and Information Retrieval

- Most techniques based on statistics over words
- No real language processing
- Processing speed a major factor for more sophisticated approaches

Technology Map: Writing Assistance

- Syntactic analysis defines the current limit of sophisticated processing
- Stylistic analysis restricted to simple pattern matching

Technology Map: Machine Translation

- Large lexica
- Rich syntactic analysis
- For transfer-based systems, structural and lexical mapping rules; limited semantic constraints
- For interlingua-based systems, some level of semantic analysis
- Current statistical systems rely on large bilingual corpora

Technology Map: Text Summarisation

- Current commercial systems use virtually no knowledge of language, other than extraction rules based on specific linguistic cues
- Statistically-based research systems now focussing on different kinds of summaries:
  - Multi-document summaries
  - Update summaries
- Interesting research direction: combination of information extraction technology with natural language generation

Technology Map: Query Systems

- Existing systems use combination of linguistic knowledge of question forms + finite state grammars
- Answers found by information retrieval with some minimal NLP
- Quality results come from string matching to hand-crafted answers for frequent questions
- Powerset has the rights to use Xerox's rich parsing technology, so some people think this could be a Google-killer …

Where The Action Is Today

- Statistical approaches to everything, but most successfully in machine translation
- Question-answering
- Voice search
- Audio mining

Finding Out More: Comprehensive Texts

- R Dale, H Moisl and H Somers (eds) [2000] Handbook of Natural Language Processing. Marcel Dekker Inc.
- D Jurafsky and J Martin [2000] Speech and Language Processing. Prentice-Hall; new edition forthcoming.
- R Cole, A Zaenen and A Zampolli (eds) [1998] Survey of the State of the Art in Human Language Technology. Cambridge University Press. Available free on the web.

Finding Out More: Research Journals

- Computational Linguistics
- Natural Language Engineering
- Machine Translation
- Speech Communication
- Computer Speech and Language

Finding Out More: Professional Associations

- Association for Computational Linguistics
  - www.aclweb.org
- European Association for Machine Translation
  - www.eamt.org
- Association for Machine Translation in the Americas
  - www.amtaweb.org/
- International Speech Communication Association
  - www.isca-speech.org

Finding Out More: Research Conferences

- Association for Computational Linguistics
- COLING: International Conference on Computational Linguistics
- Interspeech
- MT Summit

Finding Out More: Mailing Lists

- Corpora (http://nora.hd.uib.no/corpora/)
- MT-List (http://www.eamt.org/mt-list.html)
- The Linguist List (http://linguistlist.org)

Cute Applications #1: Control Your Home by Voice

- One Voice's Media Center Communicator
- www.onev.com/mcc/
- US$79.95

Cute Applications #2: The Smart Shopper

US$149.99

Cute Applications #3: The Voice Interactive Alarm Clock

Cute Applications #4: Multilingual Speech to Speech Translation

ECTACO iTRAVL NTL-9C Talking 2-way Multilingual Language Communicator and Electronic Dictionary

ECTACO iTRAVL NTL-9C Talking 2-way Multilingual Language Communicator and Electronic Dictionary

Christmas Price: £347.95

Order Now

Christmas Sale!

Buy now and get a 10-language Universal Translator ML350 (£83alue) free!

Order Now

Language pair: English <-> English, English <-> Russian, Russian <-> English, English <-> Spanish, Spanish <-> English, English <-> Polish, Polish <-> English, English <-> German, German <-> English, English <-> Portuguese, Portuguese <-> English, English <-> French, French <-> English, English <-> Italian, Italian <-> English, English <-> Chinese, Chinese <-> English,

Absolutely unique in the world of handheld language and travel assistants the multilingual iTRAVL NTL-9C includes everything the business or leisure traveler needs to make visiting foreign destinations an enjoyable and easily navigated experience.

Cute Applications #5: NEC's PaPeRo Blogbot

Follow-up Comments and Questions

- Please email rdale@ics.mq.edu.au
- Thanks for coming!