Plan for Today's Lecture(s)

- Motivating NLP
- Information Extraction
- Machine Translation
- Speech Recognition and Synthesis

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 November 2013
Lecture 26.1 – Introducing Natural Language Processing

Imagining Natural Language Processing in 1968 – "2001: A Space Odyssey"

Dave and Frank think HAL is malfunctioning, so they plan to disconnect it. But HAL reads their lips when they think they are avoiding his speech recognition…

Natural Language Processing

- NLP has the goal of creating computers and machines that can use natural language as their inputs and outputs
- The field is broad, and involves computer science, linguistics, cognitive psychology, statistics

Natural Language Processing

- Every NLP application needs to identify and classify the words, phrases, structures, documents in some context or domain
- The earliest NLP applications were very narrow in scope, relying on “hand crafted” rules to implement the required knowledge
- Today much NLP uses dictionaries, stemmers, grammars, and other language knowledge
- But the statistics of co-occurrence / conditional probability yield many practical techniques for analyzing and generating language that make no use of “languageness”

Real World NLP Applications

- Information extraction
- Summarization
- Auto-journalism (story generation)
- Machine translation
- Speech recognition and synthesis
- Computational classification
    - Topic identification
    - Author identification
    - Spam detection
    - Sentiment analysis
- Question answering, automated customer service, recommendations…

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202 “Information Organization & Retrieval” Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 November 2013
Lecture 26.2 – Information Extraction

Understanding Information Extraction

- What type of structure is being extracted?
- What is the unstructured source input?
- What resources are available to guide the extraction?
- What extraction techniques are employed?
- What is the format of the extracted information?

(IE typically starts after tokenization, part-of-speech tagging, and phrase identification steps of text processing)

"Named Entity" Recognition

- People, organizations, locations, dates, etc. can be identified with high accuracy in most kinds of documents using a combination of dictionaries, directories, gazetteers and rules – many are domain-specific
- Named entity recognition is essential for machine translation because if multi-word names are missed, translating them word-for-word will cause errors (e.g., Golden Gate Park, Walnut Creek)
- Important entities are likely to be mentioned many times in a text, but are often described by different noun phrases each time, requiring co-reference resolution

Application Areas for IE

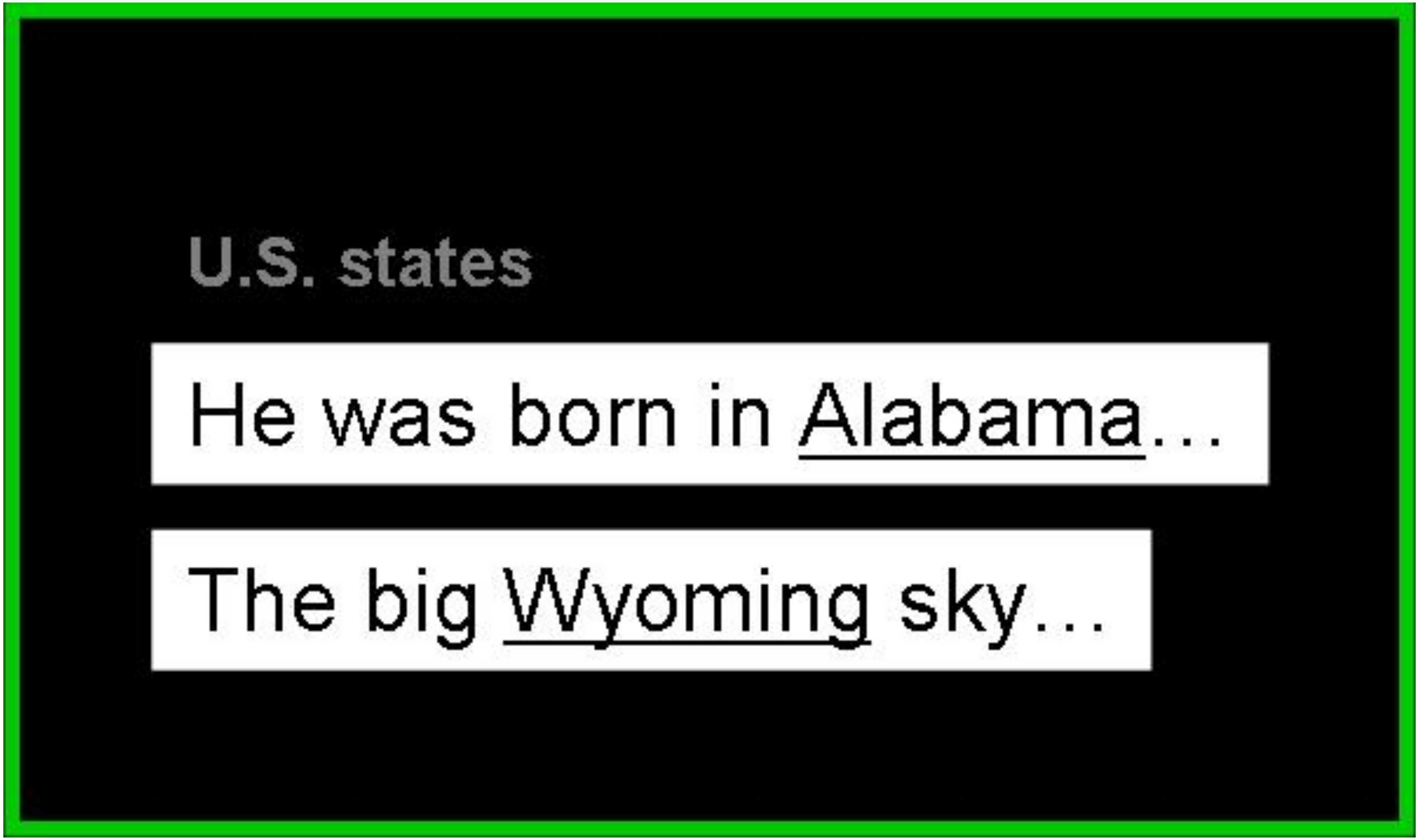
- Sales intelligence and lead generation (customers)
- Market intelligence (competitors, pricing)
- Business intelligence (aggregating information)
- "Central Intelligence" and Homeland Security

Extraction Rules

- *Capitalized-word* + "Corp." → finds company names
- "Mr." *capitalized-word+* → finds person names
- *Capitalized-word+* "," *number-below-100* "," → finds person names
- Mark Zuckerberg, 29, …
- "Mark Zuckerberg" and "Mr. Zuckerberg" are probably the same entity if these two phrases occur near each other
- But "Golden Gate Park" and "Mr. Park" aren't …

Examples from Grishman, Ralph. "Information extraction." *The Handbook of Computational Linguistics and Natural Language Processing* (2003): 515-530.

Extracting Dataypes and Patterns: Closed Sets

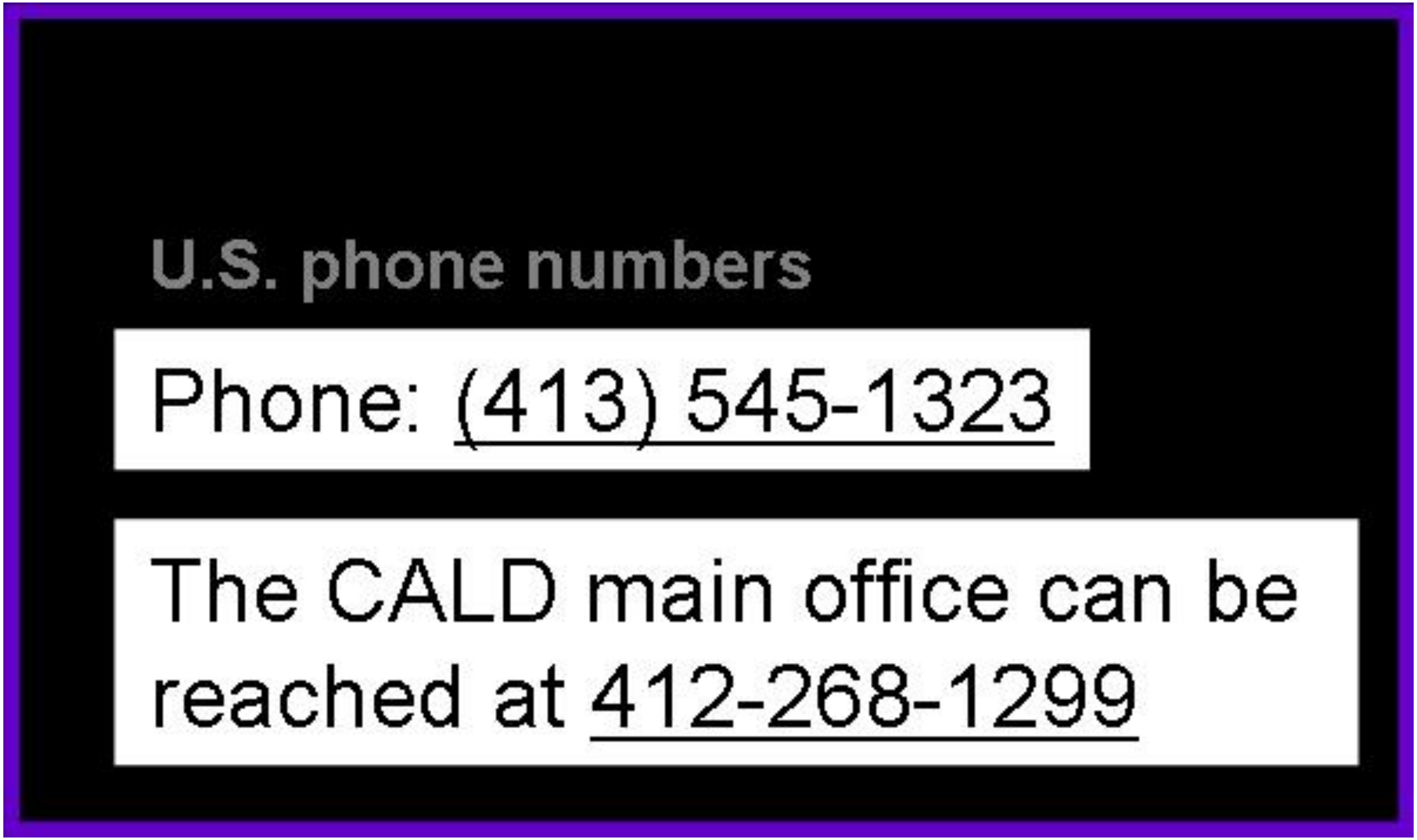


U.S. states

He was born in Alabama…

The big Wyoming sky…

Extracting Dataypes and Patterns: Regular Expressions

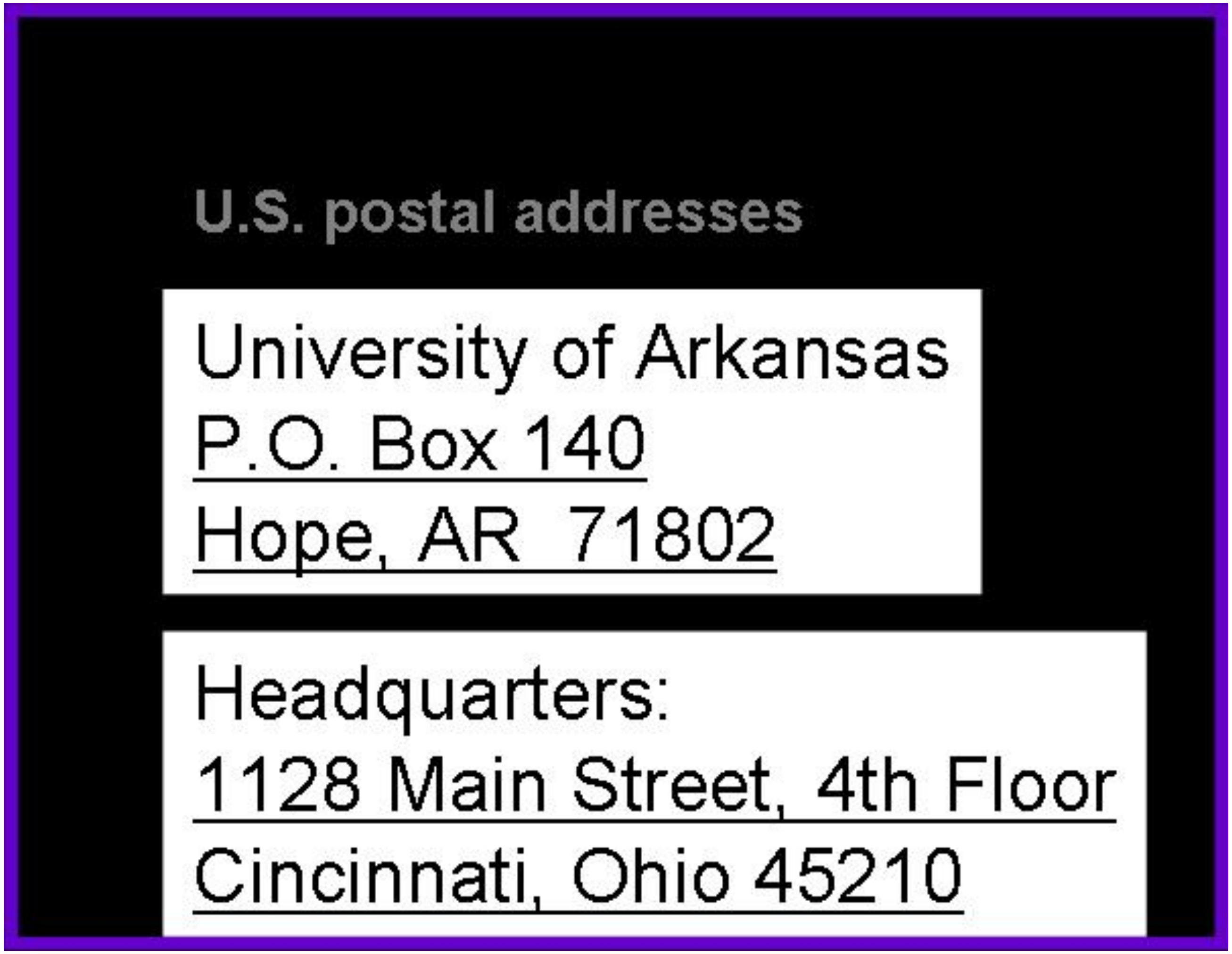


U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Extracting Dataypes and Patterns: Canonical Order



U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

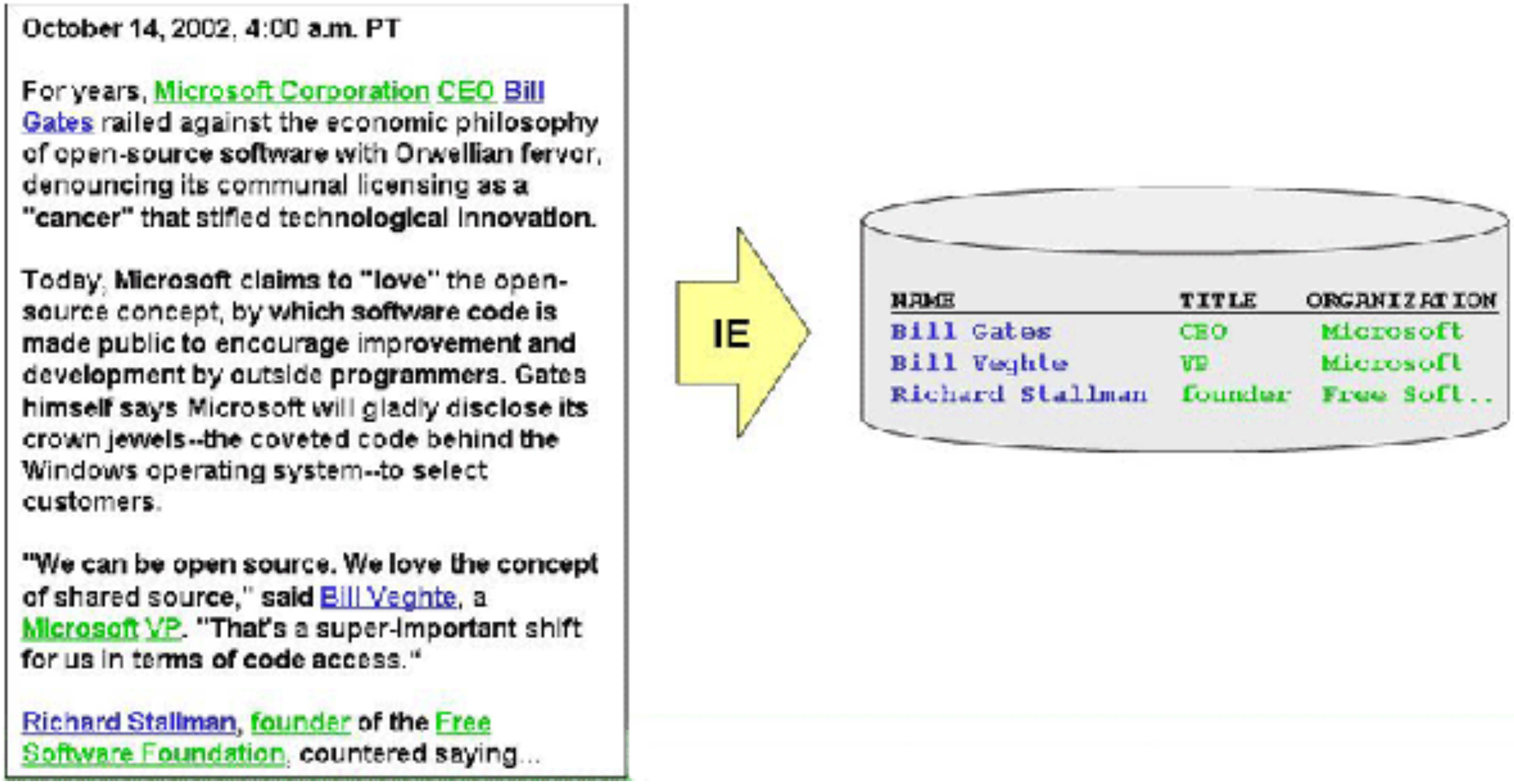
Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Extracting Multi-Way Relationships

- Often called "record extraction"
- Some N-way extractions are easy because of conventional ordering of the entities in the unstructured input (e.g., creating structured address records)
- But the goal of most N-way extractions is to populate schemas involving causal relations and dependencies (e.g., for "events" like disease outbreaks, news about company reorganizations and employee promotions)

Record Extraction

As a task: Filling slots in a database from sub-segments of text.



October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

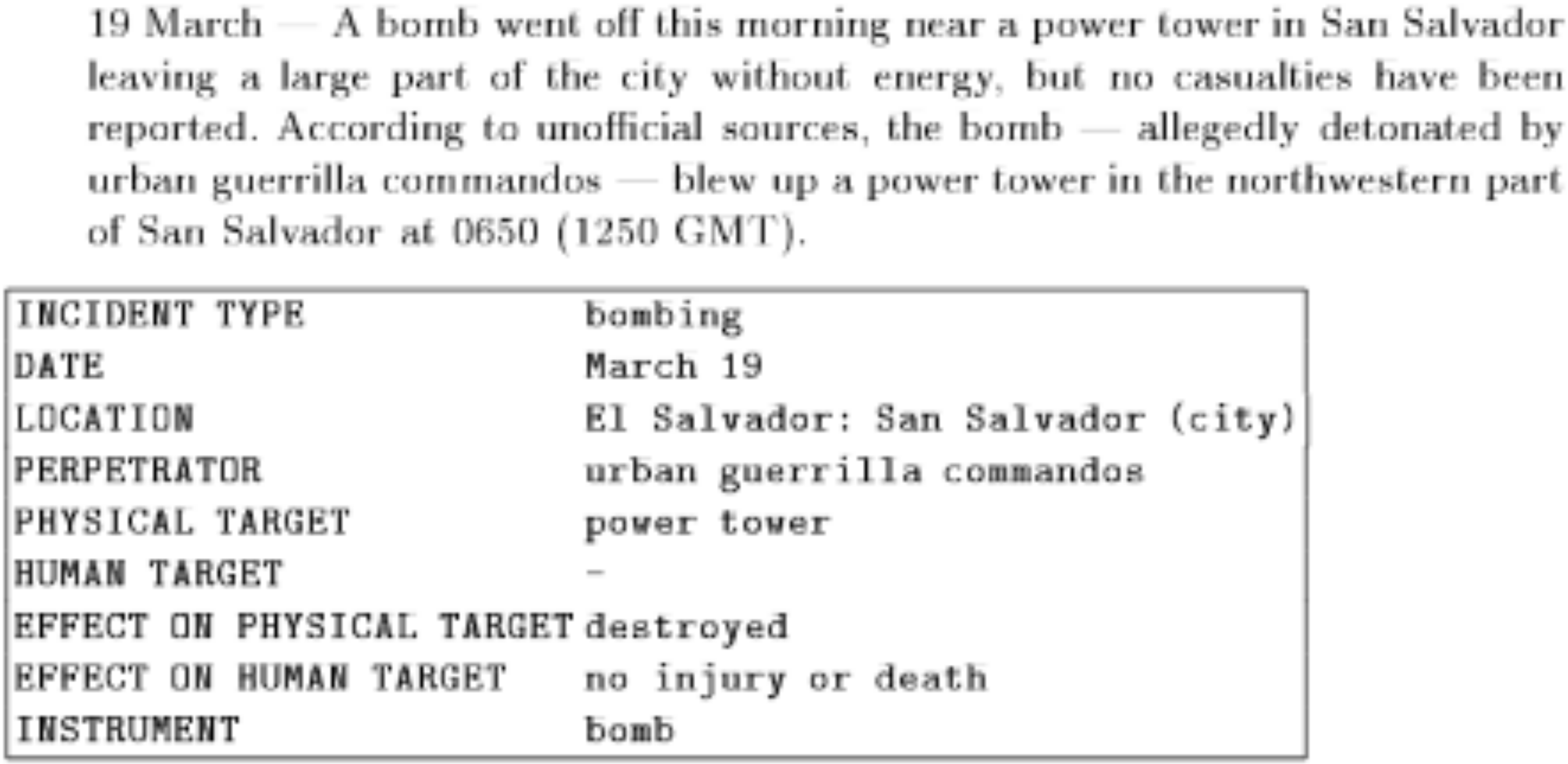
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Domain Specific Entities and Relationships



19 March — A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb — allegedly detonated by urban guerrilla commandos — blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador: San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

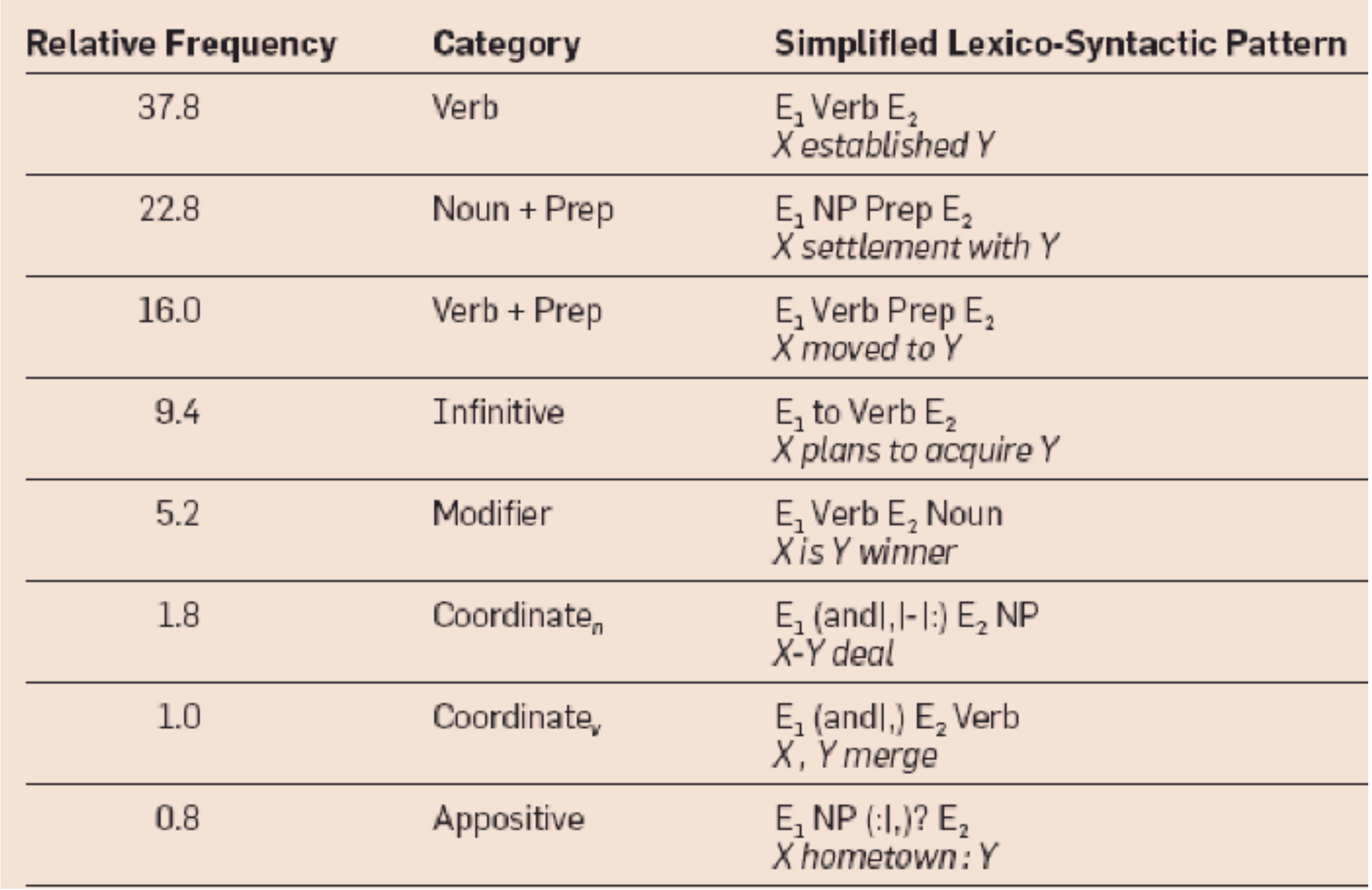
See DARPA Message Understanding Conference

“Open” Information Extraction

- The earliest information extraction applications were designed to identify a pre-specified set of entities and relationships (e.g., terrorism events)
- But these hand-crafting techniques can’t possibly scale to the web or more open-ended environments
- More recent efforts exploit the fact that most relationships follow a very small set of patterns

- Alchemy Demo

Binary Entity Relationships

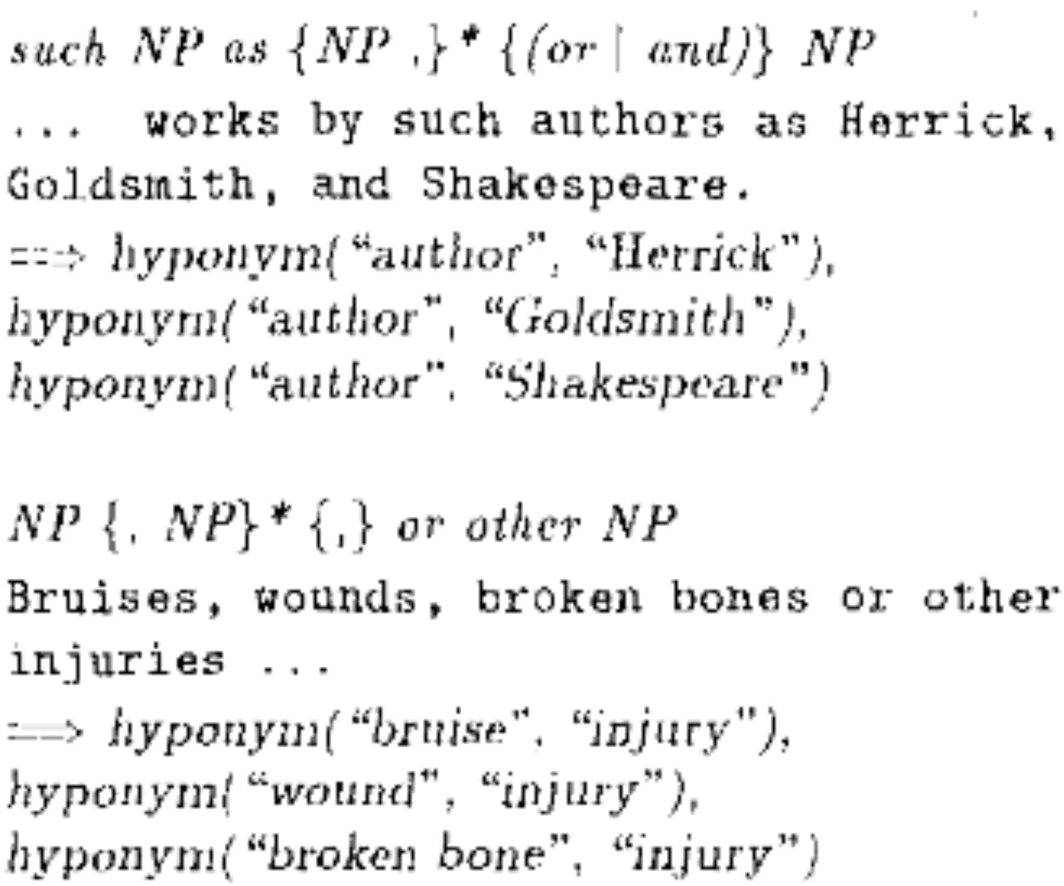


Relative Frequency	Category	Simplified Lexico-Syntactic Pattern
37.8	Verb	E_1 Verb E_2 *X established Y*
22.8	Noun + Prep	E_1 NP Prep E_2 *X settlement with Y*
16.0	Verb + Prep	E_1 Verb Prep E_2 *X moved to Y*
9.4	Infinitive	E_1 to Verb E_2 *X plans to acquire Y*
5.2	Modifier	E_1 Verb E_2 Noun *X is Y winner*
1.8	Coordinate$_n$	E_1 (and|,|-|:) E_2 NP *X-Y deal*
1.0	Coordinate$_v$	E_1 (and|,) E_2 Verb *X , Y merge*
0.8	Appositive	E_1 NP (:|,)? E_2 *X hometown : Y*

Two relationship types account for 60% of them and 8 relationship types account for 95% of them

Etzioni, Oren, et al. "Open information extraction from the web." *Communications of the ACM* 51.12 (2008): 68-74)

Identifying Hyponyms



such NP as {NP ,} {(or | and)} NP*

... works by such authors as Herrick, Goldsmith, and Shakespeare.

⟹ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

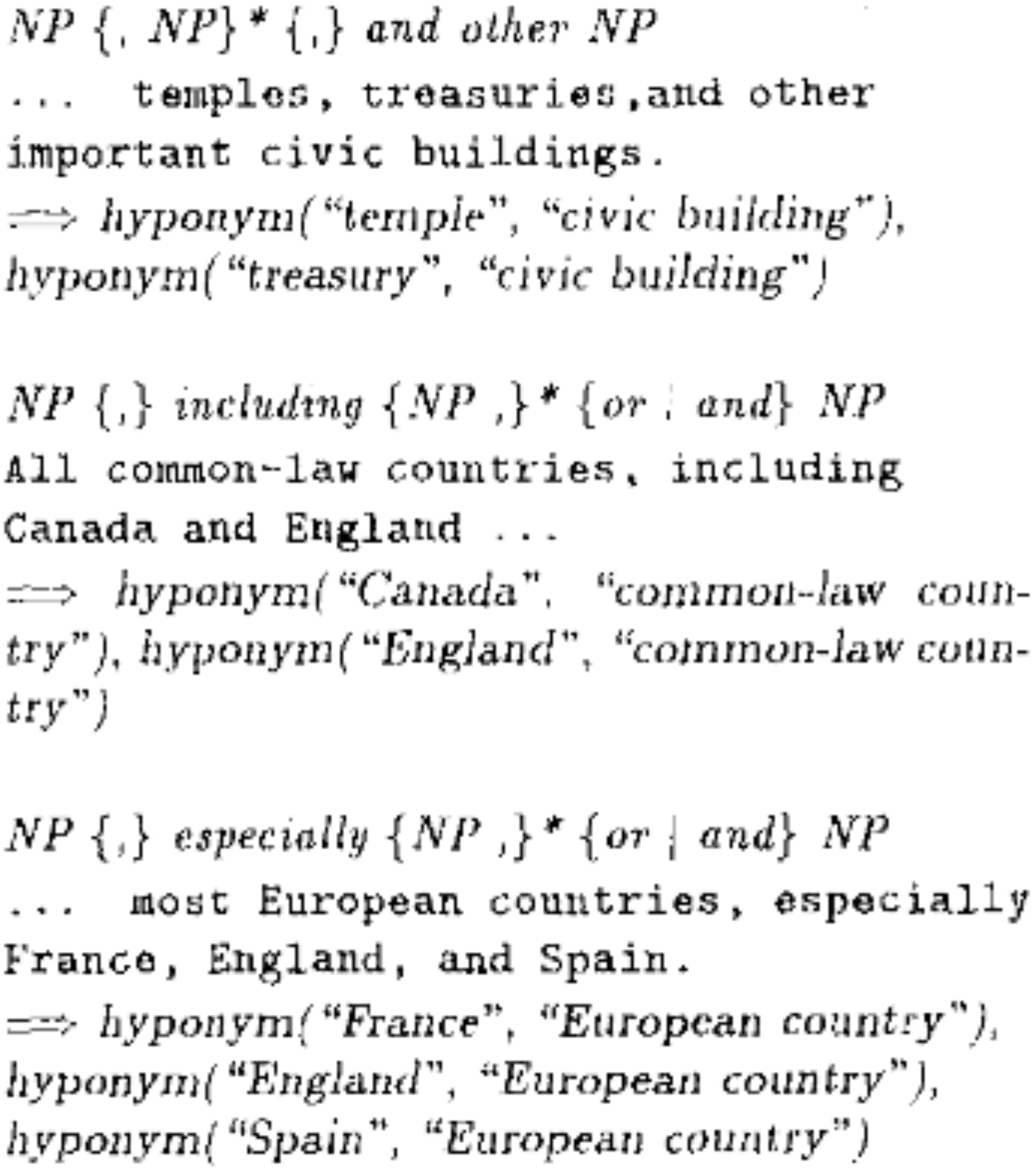
NP {, NP} {,} or other NP*

Bruises, wounds, broken bones or other injuries ...

⟹ *hyponym("bruise", "injury"),*
hyponym("wound", "injury"),
hyponym("broken bone", "injury")

Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pp. 539-545. Association for Computational Linguistics, 1992.

Identifying Hyponyms



NP {, NP}* {,} and other NP

... temples, treasuries,and other important civic buildings.

⟹ hyponym("temple", "civic building"), hyponym("treasury", "civic building")

NP {,} including {NP ,}* {or | and} NP

All common-law countries, including Canada and England ...

⟹ hyponym("Canada", "common-law country"), hyponym("England", "common-law country")

NP {,} especially {NP ,}* {or | and} NP

... most European countries, especially France, England, and Spain.

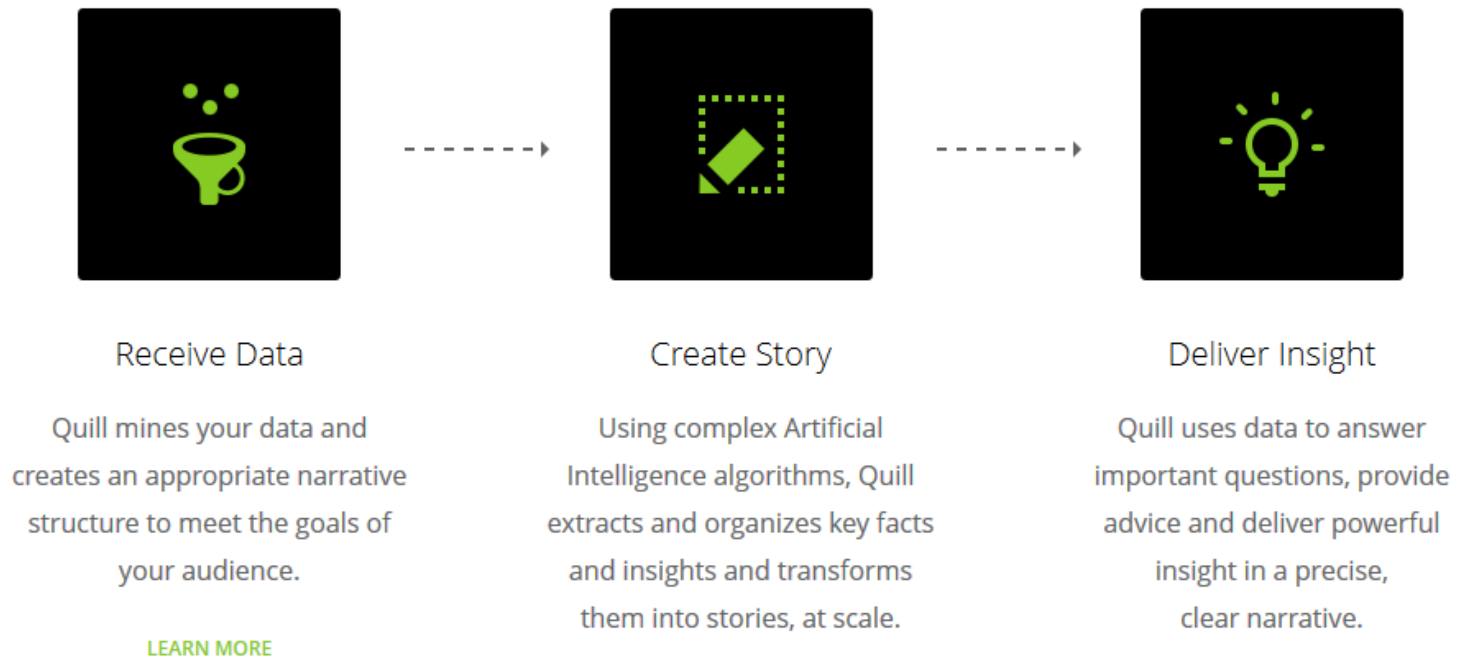
⟹ hyponym("France", "European country"), hyponym("England", "European country"), hyponym("Spain", "European country")

Text Summarization

- A summary is a text produced from one or more texts that conveys their important information(to people) using significantly fewer words
- Summarization requires *EXTRACTION* to identify important content, *ABSTRACTION* to regenerate it in more concise form, *FUSION* to combine the extracted parts, and *COMPRESSION* to eliminate unimportant content
- The "important content" is identified via a

Robo-Journalism

- Domains in which the most important information is inherently highly structured can do the inverse of information extraction



Receive Data

Quill mines your data and creates an appropriate narrative structure to meet the goals of your audience.

LEARN MORE

Create Story

Using complex Artificial Intelligence algorithms, Quill extracts and organizes key facts and insights and transforms them into stories, at scale.

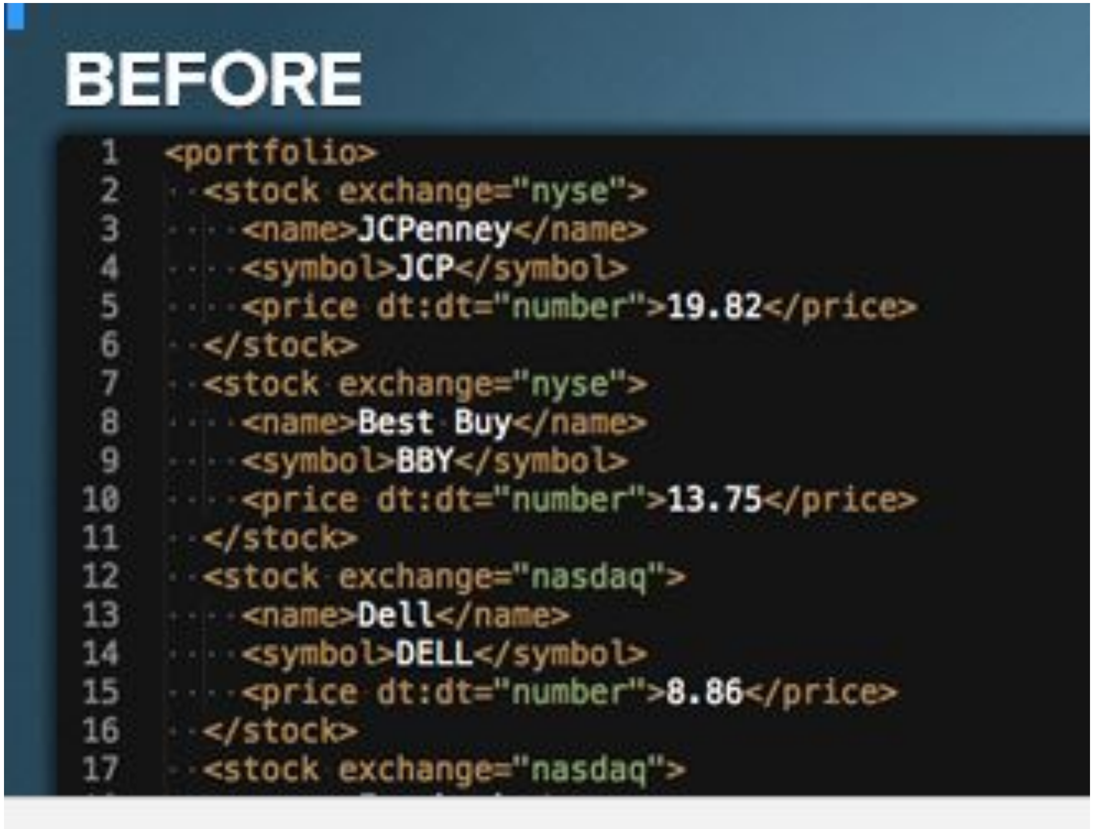
Deliver Insight

Quill uses data to answer important questions, provide advice and deliver powerful insight in a precise, clear narrative.

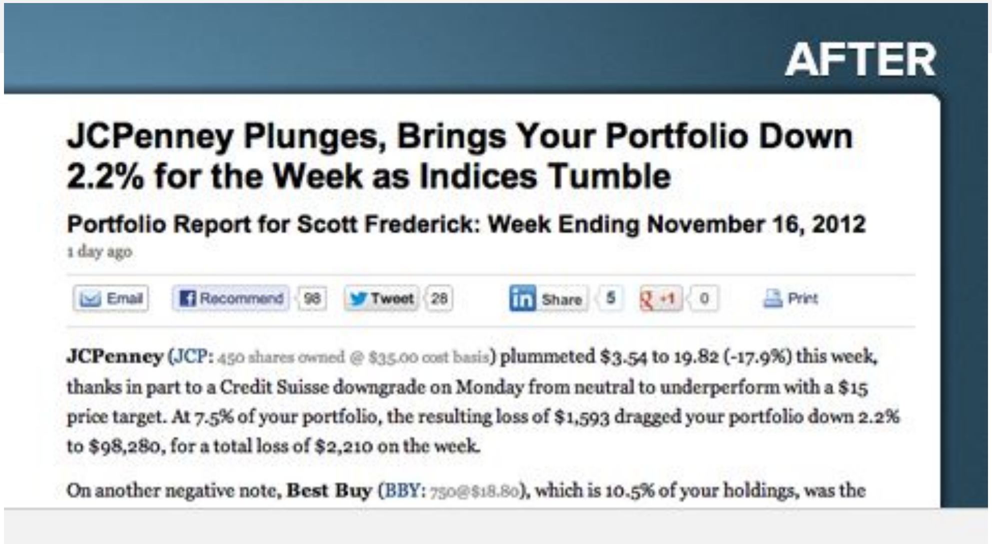
Narrative science pioneered this idea to generate sports and finance stories, which follow templates and grammars

“Automated Insights”

“Our patented artificial intelligence platform sifts through large data sets and identifies key patterns and trends, then describes those insights in plain English with the tone, personality and variability of a human writer”



```
<portfolio>
  <stock exchange="nyse">
    <name>JCPenney</name>
    <symbol>JCP</symbol>
    <price dt:dt="number">19.82</price>
  </stock>
  <stock exchange="nyse">
    <name>Best Buy</name>
    <symbol>BBY</symbol>
    <price dt:dt="number">13.75</price>
  </stock>
  <stock exchange="nasdaq">
    <name>Dell</name>
    <symbol>DELL</symbol>
    <price dt:dt="number">8.86</price>
  </stock>
  <stock exchange="nasdaq">
```



JCPenney Plunges, Brings Your Portfolio Down 2.2% for the Week as Indices Tumble

Portfolio Report for Scott Frederick: Week Ending November 16, 2012

1 day ago

JCPenney (JCP: 450 shares owned @ $35.00 cost basis) plummeted $3.54 to 19.82 (-17.9%) this week, thanks in part to a Credit Suisse downgrade on Monday from neutral to underperform with a $15 price target. At 7.5% of your portfolio, the resulting loss of $1,593 dragged your portfolio down 2.2% to $98,280, for a total loss of $2,210 on the week.

On another negative note, **Best Buy** (BBY: 750@$18.80), which is 10.5% of your holdings, was the

Story Templates and Grammars

- Some document types have a regular "discourse" or "frame" structure that makes it easier to find information that fills the key "slots“
  - What are the slots in a generic sports story? What if the sports are subdivided?
  - Election story? “Weather event” story? Etc….

***Once Narrative Science** had mastered the art of telling sports and finance stories, the company realized that it could produce much more than journalism. Indeed, anyone who needed to translate and explain large sets of data could benefit from its services…*

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 November 2013
Lecture 26.3 – Machine Translation

Machine Translation: An Apocryphal but Important Example

- A story often told about the early days of machine translation research (1950s) is that the English sentence:

The spirit is willing, but the flesh is weak

when translated into Russian, and then back to English became:

The vodka is strong but the meat is rotten

Machine Translation: A Brief History [1]

- Initial optimism in the 1950s was followed by extreme pessimism
- In 1966 the Automatic Language Processing Advisory Committee (ALPAC) concluded "there is no immediate or predictable prospect of useful machine translation" and recommended the development of computer aids for human translators
- Fortunately, ALPAC also recommended continued support of basic research in computational linguistics

Machine Translation: A Brief History [2]

- In the 1970s and 1980s MT systems continued to develop; the dominant technical strategy relied on hand-crafted syntactic parsers, morphological analyzers, and dictionaries - intensely semantic and rule-based approaches.
- The 1990s were a major turning point. IBM research developed the Candide system that relied purely on statistical analysis and "example-based" methods for phrase matching and translation

Machine Translation: A Brief History [3]

- Candide used a very large corpus of English and French documents that had extremely reliable bi-directional translations
- This bilingual corpus contained enough examples to estimate the substitutability or semantic equivalence of words between English and French

Using Statistics to Improve Translation

Word selection in translation:

French phrase *groupe de travail*
groupe translates to cluster, group, grouping, concern, collective
travail translates to work, labor, labour

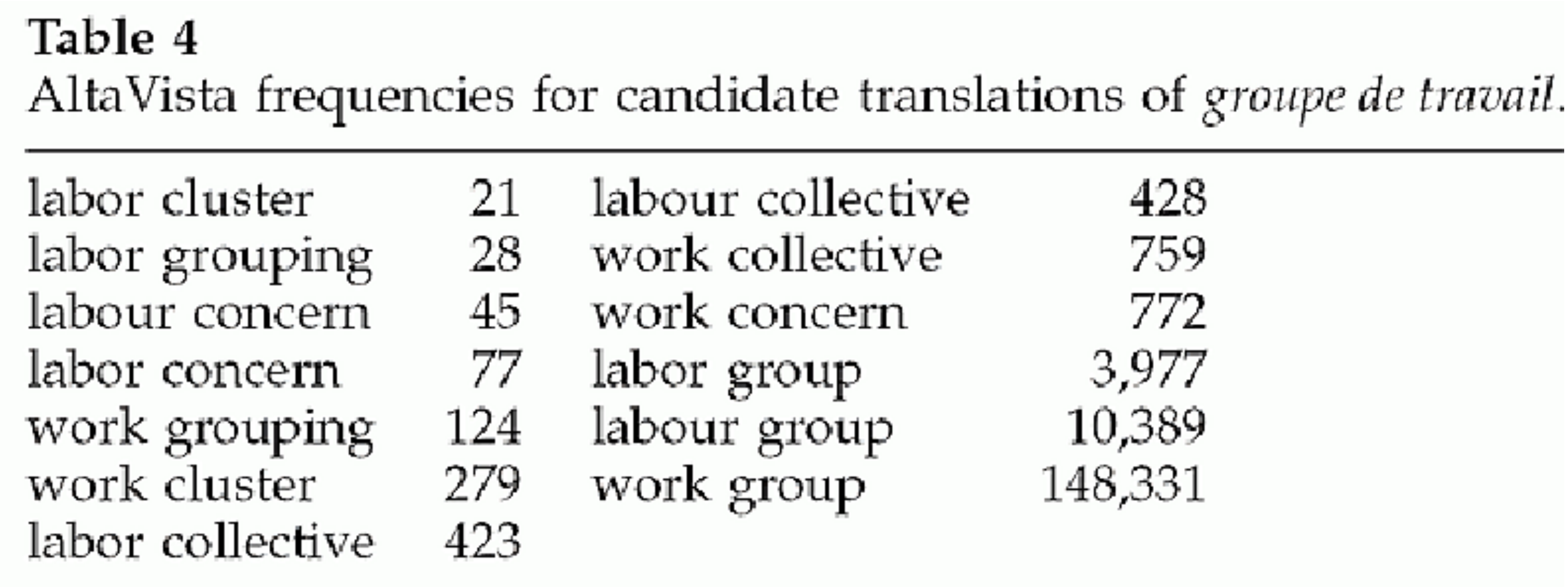
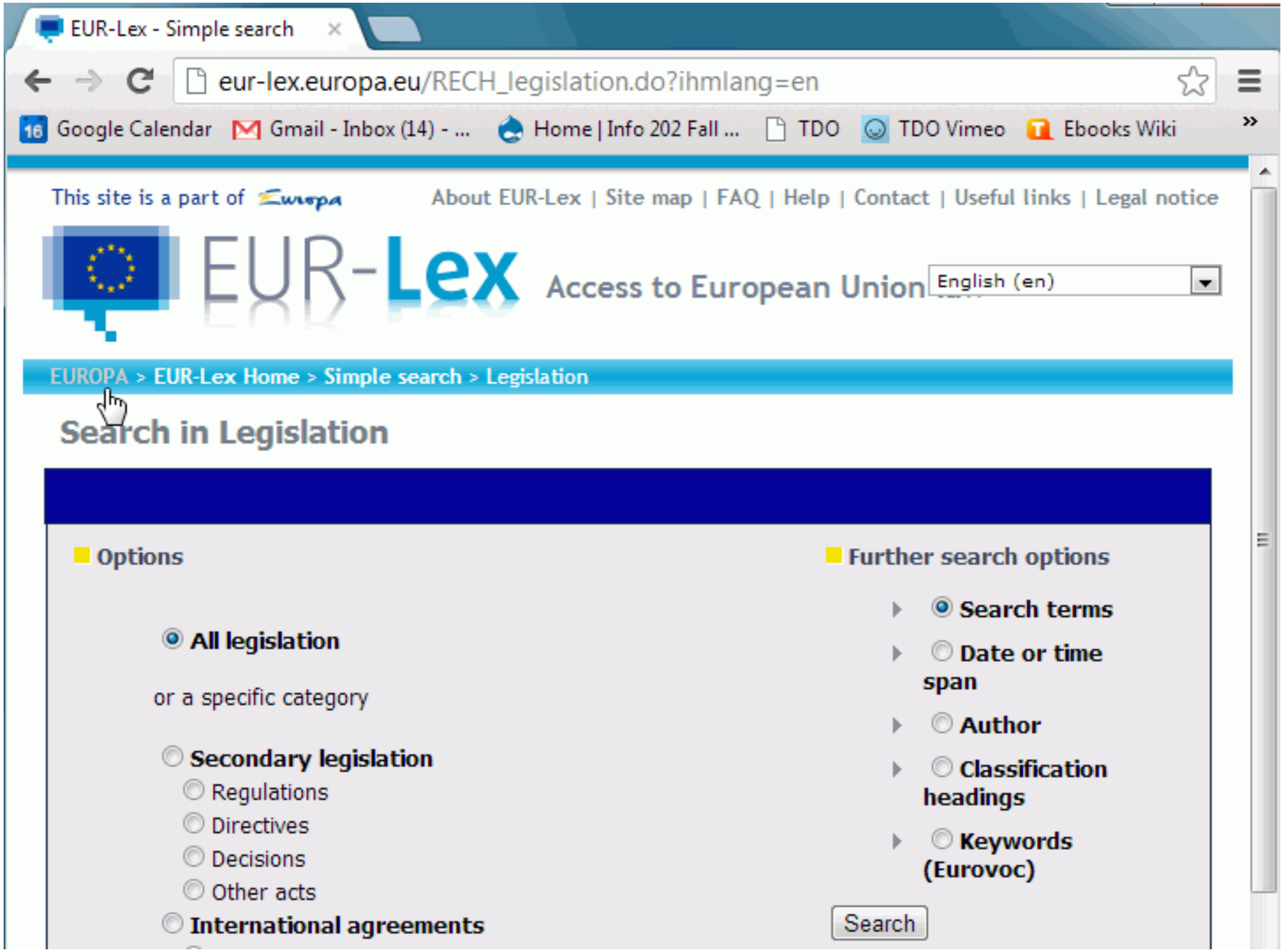


Table 4
AltaVista frequencies for candidate translations of *groupe de travail*.

labor cluster	21	labour collective	428
labor grouping	28	work collective	759
labour concern	45	work concern	772
labor concern	77	labor group	3,977
work grouping	124	labour group	10,389
work cluster	279	work group	148,331
labor collective	423		

Authoritative Multi-Language Corpus


EUR-Lex - Simple search
eur-lex.europa.eu/RECH_legislation.do?ihmlang=en
Google Calendar
Gmail - Inbox (14) - ...
Home | Info 202 Fall ...
TDO
TDO Vimeo
Ebooks Wiki
This site is a part of Europa
About EUR-Lex | Site map | FAQ | Help | Contact | Useful links | Legal notice
EUR-Lex
Access to European Union law
English (en)
EUROPA > EUR-Lex Home > Simple search > Legislation
Search

Search in Legislation

Options

● All legislation

or a specific category

○ Secondary legislation

○ Regulations

○ Directives

○ Decisions

○ Other acts

○ International agreements

Further search options

● Search terms

○ Date or time span

○ Author

○ Classification headings

○ Keywords (Eurovoc)

- The European Union keeps official documents from all EU institutions, including minutes of parliament hearings, all European Commission documents, regulations in every language

Authoritative Multi-Language Translation

- Today the largest corpus of authoritative multi-language translations is at the European Parliament, where there are 23 "official languages“ and realtime translation is often necessary but not always possible
- How many pairwise translations might need to be done?
- Is this an effective method of translation? Why or why not?

Text Corpora

- Computational linguists, computer scientists, experimental psychologists and others rely on text corpora for their research
- Prominent pre-web examples include the Brown corpus (Kucera and Francis, 1967) that includes a million words of contemporary American English
- The web dwarfs any other (possible?) corpus -- Google probably indexes a few trillion words, making it orders of magnitude larger than any other text collection (with the possible exception of Microsoft's)

Pre-Nominal Adjective Ordering

- In translation and text generation we need to arrange nouns and adjectives to "sound right" and create the intended meaning:
- *Big brown dog* is grammatical
- *Brown big dog* is not
- This is a challenging problem for ESL people, and especially when their native language does it very differently

Getting Adjective Order Right

- Some NLP approaches attempt this using rules (e.g., size adjectives precede color ones)
- But it can also be done using data-intensive approaches; compare frequency of {a, b} with {b, a}
- Generate all the possible strings and choose the one that occurs most frequently
- "smart beautiful woman" 5.5 x more frequent than "beautiful smart woman" (Google search on 29 November 2010)

How Good is Machine Translation?

Microsoft's release of its Xbox 360 video-games console begins a new phase in the battle to remove Sony's PlayStation from the top spot. If it succeeds, the software giant may be tempted to make more incursions into the competitive market for home-entertainment hardware.

How Good is Machine Translation – “Google Translate” Roundtripping (German)

- (Nov 2005) Release Microsofts of its video game console Xbox 360 begins a new phase in the battle for removing from PlayStation Sonys from the upper point. If it follows, the software giant can be provoked, in order to form more ideas into the free market for house maintenance small articles.
- (Nov 2013) Microsoft's release of the Xbox 360 video game console begins a new phase in the struggle for Sony's PlayStation remove from the top. If it is successful, the software giant may be tempted to make more inroads in the competitive market for home entertainment hardware.

How Good is Machine Translation – “Google Translate” Roundtripping (Chinese)

- (Nov 2005) Its Xbox 360 video-games control bench Microsoft. The s release starts one new stage removes Sony in this battle; s PlayStation from this top spot. If it succeeds, perhaps the software giant does invades into the competitive market for the family entertainment hardware
- (Nov 2013) Its Xbox360 video game console, Microsoft released the beginning of a new phase in the battle of Sony's PlayStation deleted from the top spot. If successful, the software giant might be tempted to make more broke into a competitive market, home entertainment hardware.

Is Machine Translation Good Enough? Compared to What?

- Is "round tripping" a fair test of machine translation? What exactly is it testing?
- English is (one of) the most widely spoken and written languages, and is the language most widely learned as a second or third one
- How well would speakers and readers of second languages compare to automated language translation?
- Automated translation is clearly more accurate than some large portion of them would be

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

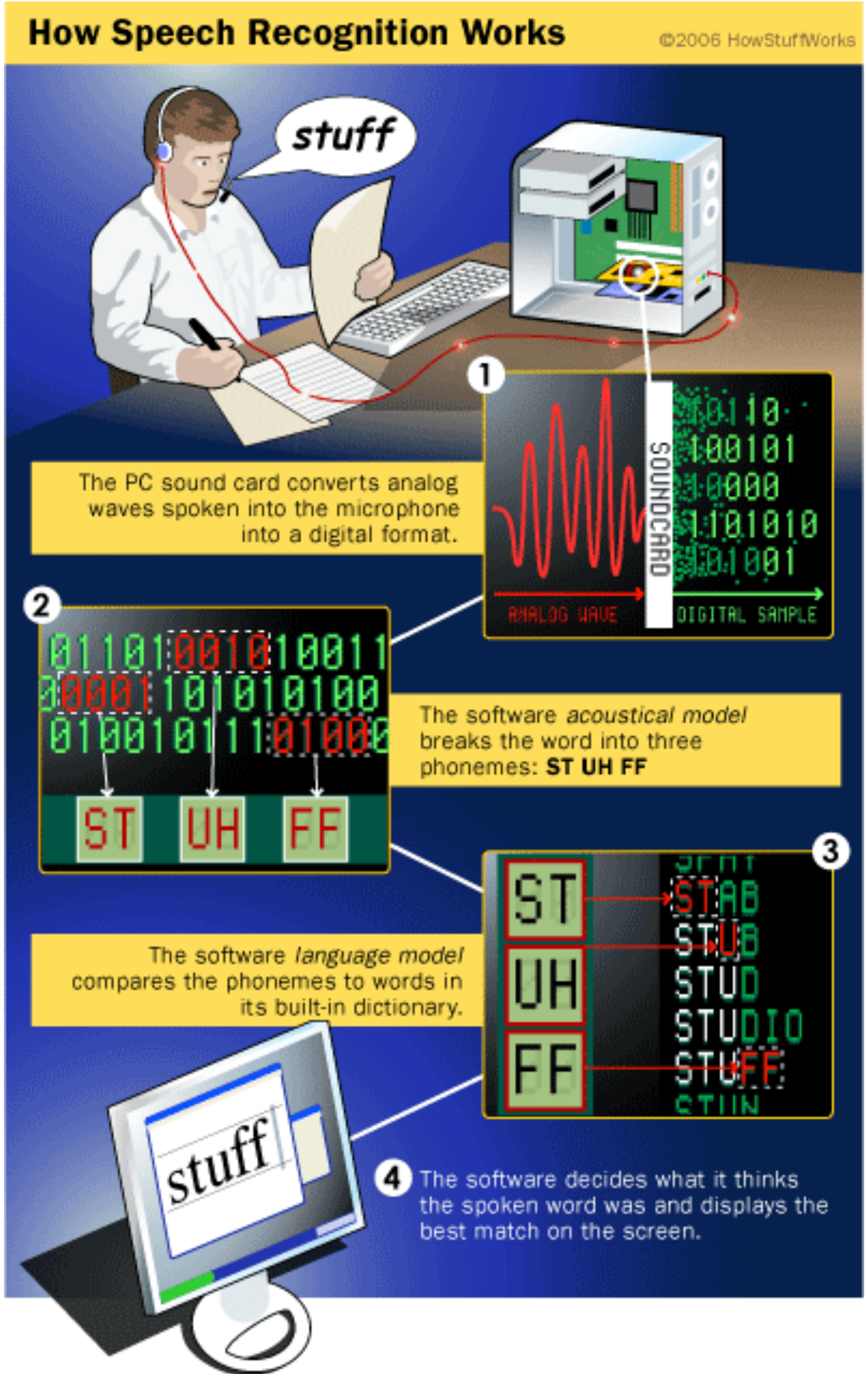
26 November 2013
Lecture 26.4 – Speech Recognition and Synthesis

Speech Recognition

- As we saw in the *2001: A Space Odyssey* example speech recognition has great potential to enhance how people interact with machines
- Like machine translation, speech recognition has a long history in NLP, but it raises many additional technical challenges because of the acoustical and signal processing challenges
- And like machine translation, speech processing had a rule-based phase but it is now much more statistical in character

Speech Recognition - Simplified Process

(about 2008)



How Speech Recognition Works

©2006 HowStuffWorks

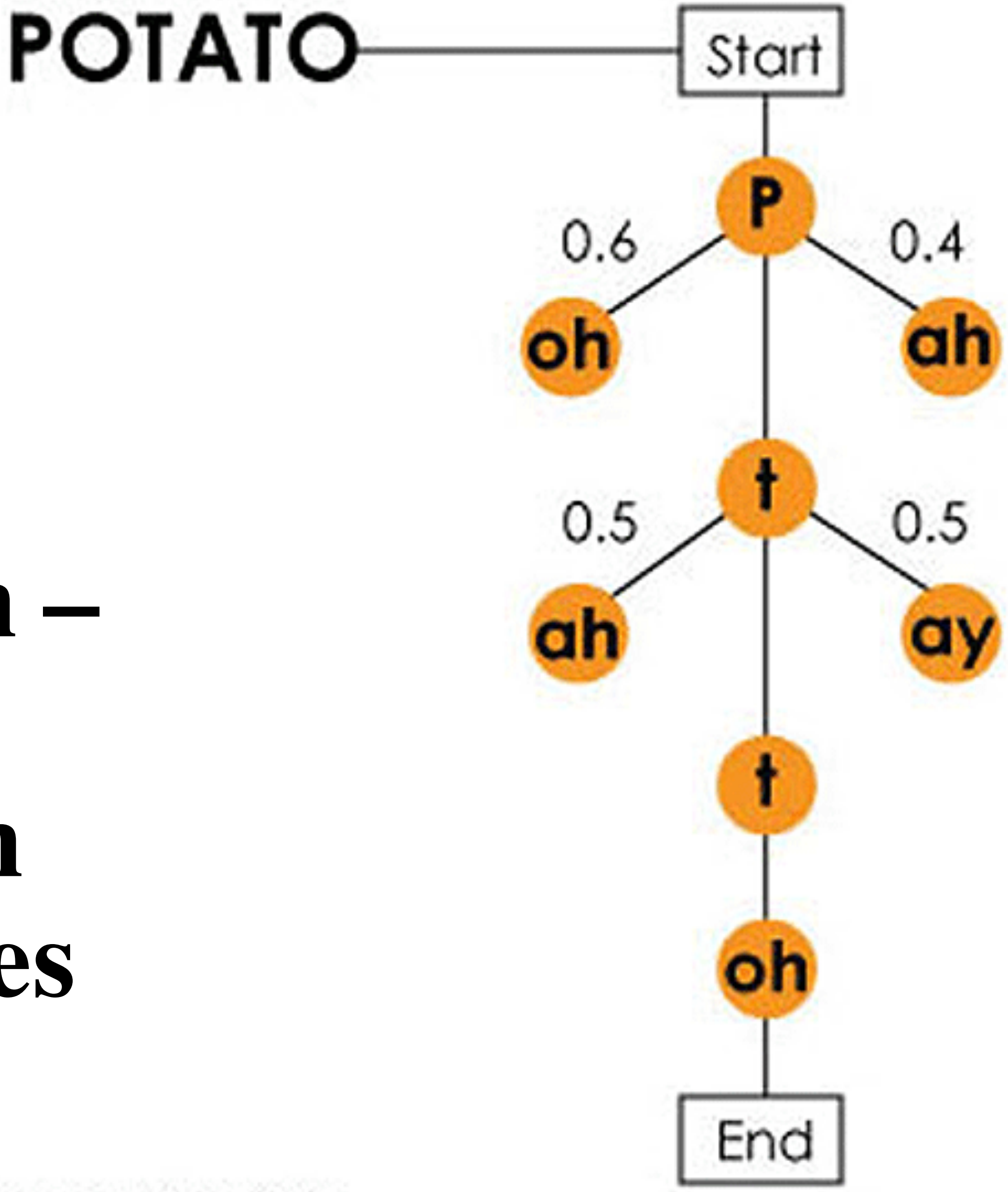
1 The PC sound card converts analog waves spoken into the microphone into a digital format.

2 The software *acoustical model* breaks the word into three phonemes: **ST UH FF**

3 The software *language model* compares the phonemes to words in its built-in dictionary.

4 The software decides what it thinks the spoken word was and displays the best match on the screen.

Speech Recognition – Using Transition Probabilities



Markov Model

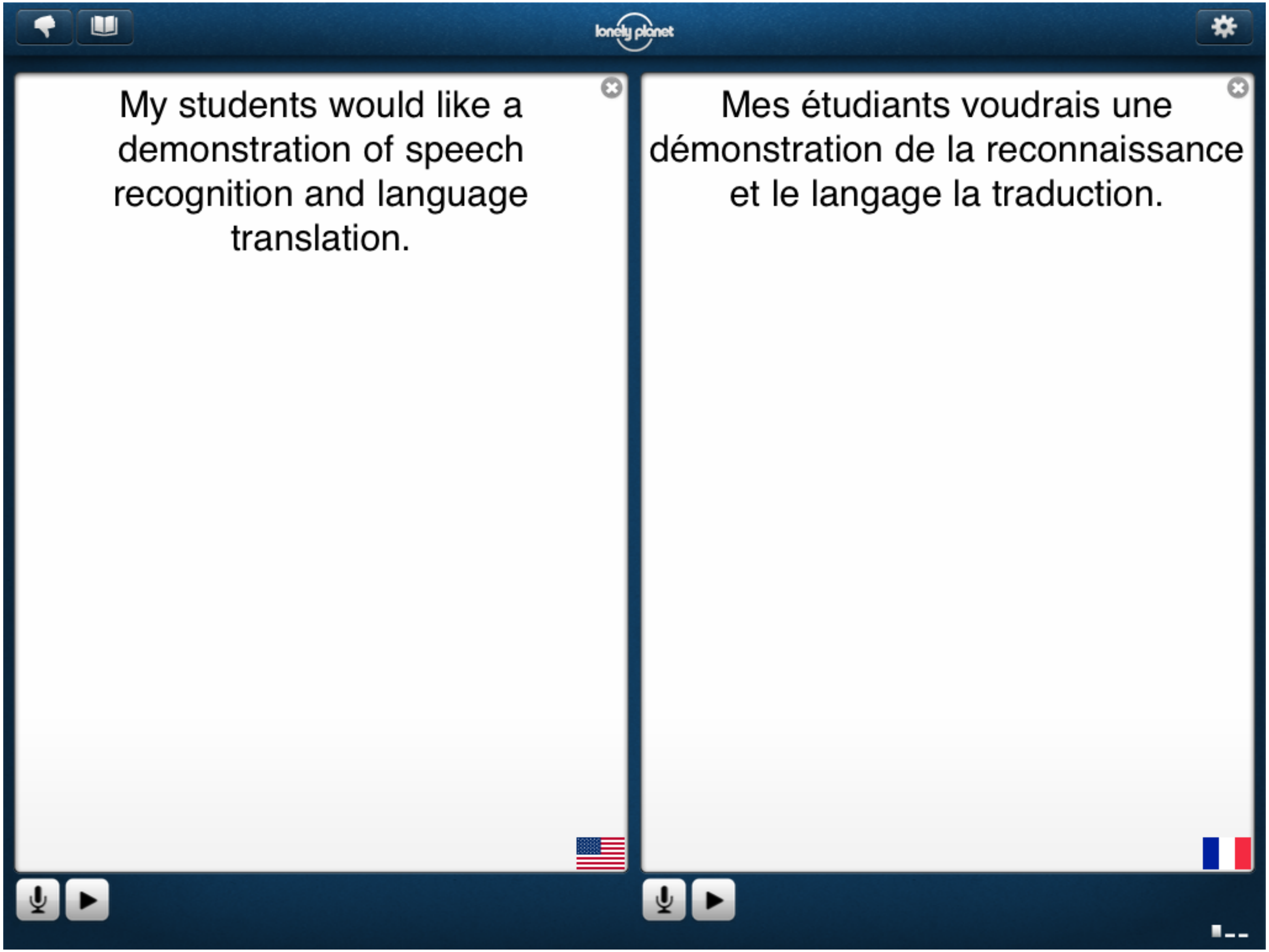
Speech Recognition – for "Consumers"

- Continuous speech recognition systems for dictation can be extremely accurate if the system has been trained to recognize the speaker's voice, but even untrained systems are getting really good (e.g., Dragon from Nuance)
- SIRI does its speech recognition "in the cloud" and emphasizes entity and phrase recognition rather than sentence and intent understanding

Speech Synthesis (Text-to-Speech)

- TTS seems like the reverse of speech recognition, but there is limited technology overlap and neither involves much language understanding
  - Being able to synthesize speech from computer-readable text is useful in applications that require spontaneous interaction ...
  - ... or where reading isn't practical, like when you're driving and need directions.
  - Speech synthesis enables accessibility for people who are sight- or voice-impaired

Speech Recognition, Translation, and Synthesis Combined



My students would like a demonstration of speech recognition and language translation.

Mes étudiants voudrais une démonstration de la reconnaissance et le langage la traduction.

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 November 2013
Lecture 26.4 – Assignment

Your Assignment: Enjoy Thanksgiving