

# Linguistic Tools for Advanced Information Extraction

Aaron Kaplan  
XRCE Parsing & Semantics group

Frédérique Segond (Manager)  
Salah Aït-Mokhtar  
Caroline Brun  
Maud Ehrmann  
Vianney Grassaud

Caroline Hagège  
Gilbert Rondeau  
Claude Roux  
Ágnes Sándor  
Anne Schiller

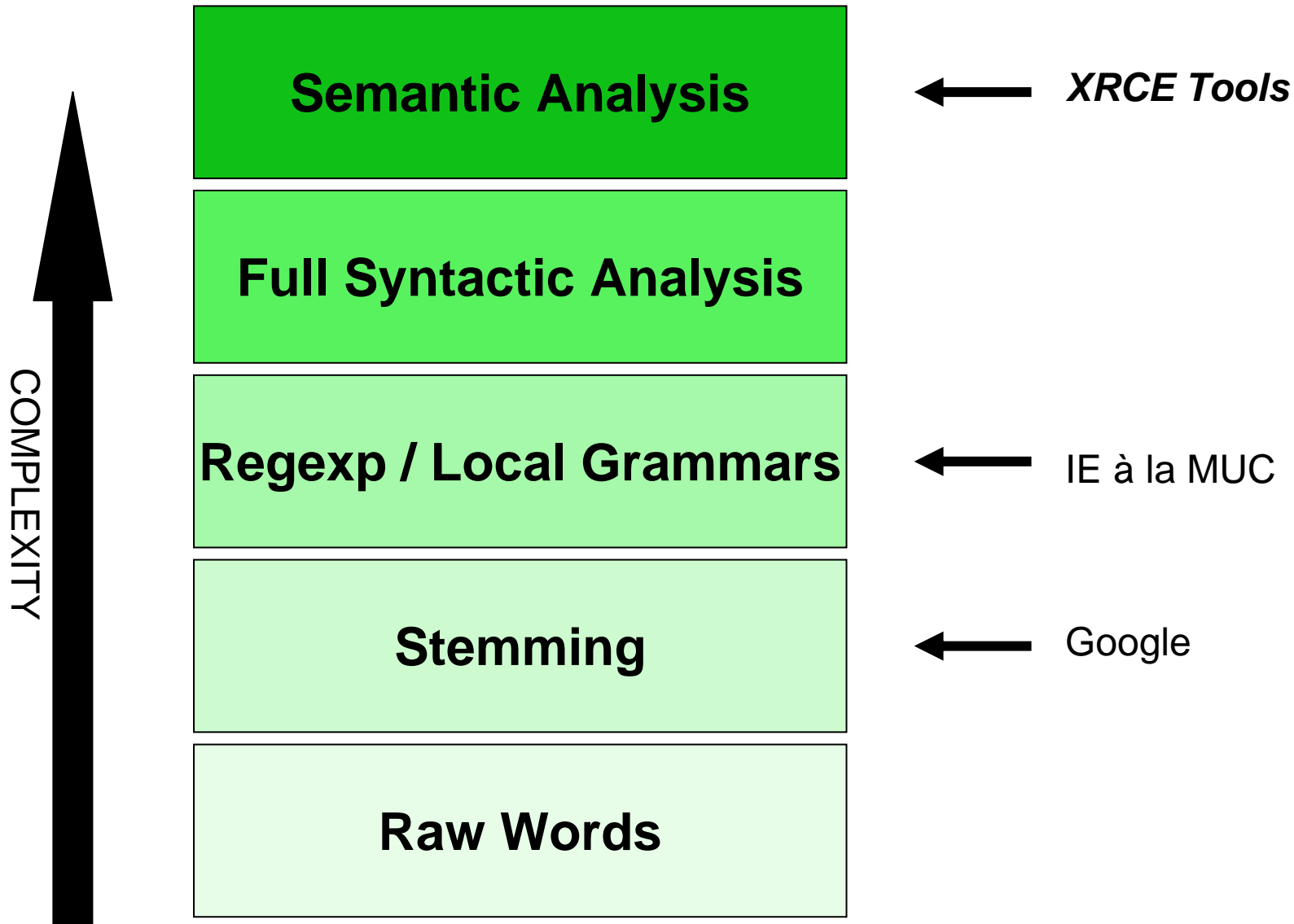
# Fact Extraction: From Finding Documents to Grasping Statements

*Information Retrieval:* finding documents on a given topic, or documents that might answer a given question

*Fact Extraction:* finding specific types of information within a set of documents, e.g.

- Construct a list of all people mentioned in a document.
- Find descriptions of mergers and acquisitions; identify who bought whom, and for how much.
- Build a timeline of events from a biographical sketch.
- Identify risks associated with medications.

# Hierarchy of Information Discovery Methods



# Xerox Incremental Parser (XIP)

The XIP engine is the core of XRCE's suite of tools

- Fast
- Robust; broad coverage
- Adaptable to multiple languages
- XML-aware

# Incremental Syntactic Analysis

<i>Hervé</i>	<i>Gallaire</i>	<i>joined</i>	<i>Xerox</i>	<i>in</i>	<i>1992</i>
Hervé Noun Proper	Gallaire Noun Proper	join Verb Past 123SP	xerox Verb Pres Non3sg	in Prep  in Adv	1992 Dig Year
		joined Adj	Xerox Noun Proper Bus		

- Tokenization & morphological analysis (FST)

# Incremental Syntactic Analysis

<i>Hervé</i>	<i>Gallaire</i>	<i>joined</i>	<i>Xerox</i>	<i>in</i>	<i>1992</i>
Hervé Noun Proper	Gallaire Noun Proper	join Verb Past 123SP	Xerox Noun Proper Bus	in Prep	1992 Dig Year

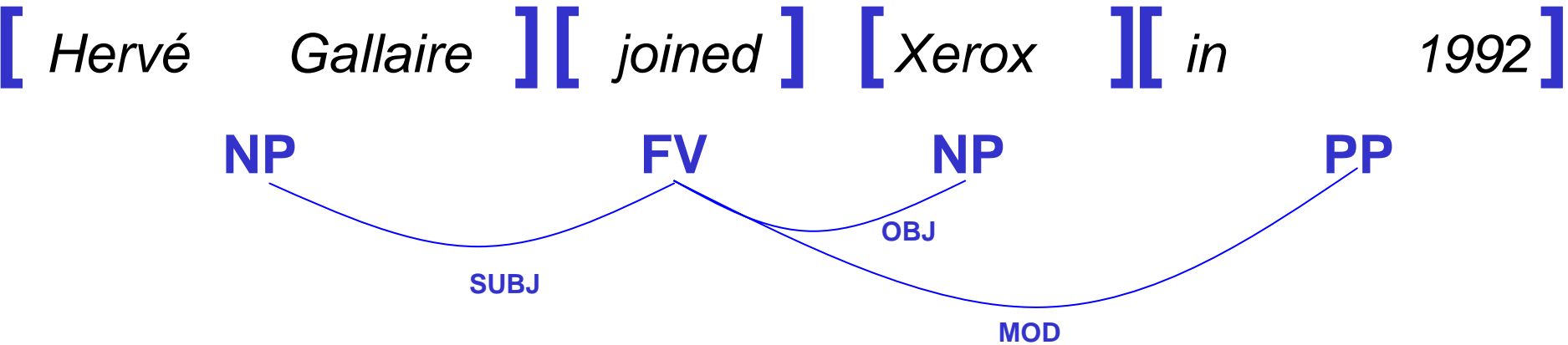
- Tokenization & morphological analysis (FST)
- Part of speech disambiguation (HMM/rules)

# Incremental Syntactic Analysis

[ *Hervé*   *Gallaire* ] [ *joined* ] [ *Xerox* ] [ *in*            *1992* ]

- Tokenization & morphological analysis (FST)
- Part of speech disambiguation (HMM/rules)
- Chunking (local rules)

# Incremental Syntactic Analysis



- Tokenization & morphological analysis (FST)
- Part of speech disambiguation (HMM/rules)
- Chunking (local rules)
- Dependency extraction (long-distance rules)



# Semantic Analysis: Named Entities

person

date

location

organization

**Anne M. Mulcahy** is chairman of the board and chief executive officer of **Xerox Corporation**, **Stamford, Conn.** She was named CEO of **Xerox** on **August 1, 2001** and chairman on **Jan. 1, 2002**. **Mulcahy** most recently was president and chief operating officer of **Xerox** from **May 2000** through **July 2001**. Prior to that, she was president of **Xerox's General Markets Operations**, which created and sold products for reseller, dealer and retail channels.

**Mulcahy** became chief staff officer in **1997** and corporate senior vice president in **1998**. Prior to that, she served as vice president and staff officer for **Customer Operations**, covering **South America** and **Central America**, **Europe**, **Asia** and **Africa** and **China**.

**Mulcahy** earned a bachelor of arts degree in **English/Journalism** from **Marvmount College** in **Tarrytown, N.Y.** in **1974**. In addition to the **Xerox board**, she is a member of the boards of directors of **Fuji Xerox Company, Ltd.**, **Target Corporation**, **Catalyst** and **Fannie Mae**, and is a member of **The Business Council**.

# Semantic Analysis: Normalization

A step towards semantic representation by obtaining more abstract labeled relations between textual elements.

Different syntactic structures can map to same normal form:

*The appointment of Gallaire as CTO...*  
*Gallaire was appointed CTO.*  
*Mulcahy appointed Gallaire CTO*



PATIENT(appoint, Gallaire)  
RESULT(appoint, CTO)

The same syntactic structure can map to different normal forms:

*He worked **for** money*

*A train **for** London*

*He has been waiting **for** hours*

PURPOSE

DESTINATION

DURATION

# Semantic Analysis: Temporal Expressions

*Hervé J. Gallaire is president of the Xerox Innovation Group and Xerox's chief technology officer. He was appointed to this position in October 2001 and was named a corporate senior vice president in December 1999. ... He joined Xerox in 1992 and managed the Xerox Research Centre Europe in Grenoble, France, from 1993 to 1998.*

- **join(Hervé J. Gallaire, Xerox), time(in 1992)**
- **manage(Hervé J. Gallaire, Xerox research Centre Europe), time(from 1993 to 1998)**
- **appointed(Hervé J. Gallaire, CTO), time(in October 2001)**

Temporal expressions (not only dates) are detected and associated with events; events are chronologically ordered independently of their appearance order in the document.

# Semantic Analysis: Coreference

Pronominal cataphora link to named entity:  
*President George W. Bush*<sub>7</sub>

Pronominal anaphora links to named entity:  
*President George W. Bush*<sub>7</sub>

Author tag for quotation:  
*John Bolton*<sub>4</sub>

Nominal coreference link to named entity:  
*President George W. Bush*<sub>7</sub>

By choosing **Paul Wolfowitz**<sub>0</sub> for the post of **World Bank**<sub>2</sub> president right after he<sub>7</sub> nominated **John Bolton**<sub>4</sub> **US**<sub>5</sub> ambassador to **the United Nations**<sub>6</sub> , **President George W. Bush**<sub>7</sub> has signaled his<sub>7</sub> determination to send his<sub>7</sub> administration 's hardliners to the forefront of the international arena.

...

Despite his<sub>0</sub> assurances , Wolfowitz<sub>0</sub> does not come off as a specialist on poverty and international development issues .

...

For his<sub>4</sub> part , Bolton<sub>4</sub> who will defend the US<sub>5</sub> administration 's foreign policy at the United Nations<sub>6</sub> , has at times in the past been tough on the world body.

*"There is no such thing as the United Nations<sub>6</sub>," he<sub>4</sub> stated in 1994.*

...

**Quotation: John Bolton**<sub>4</sub>

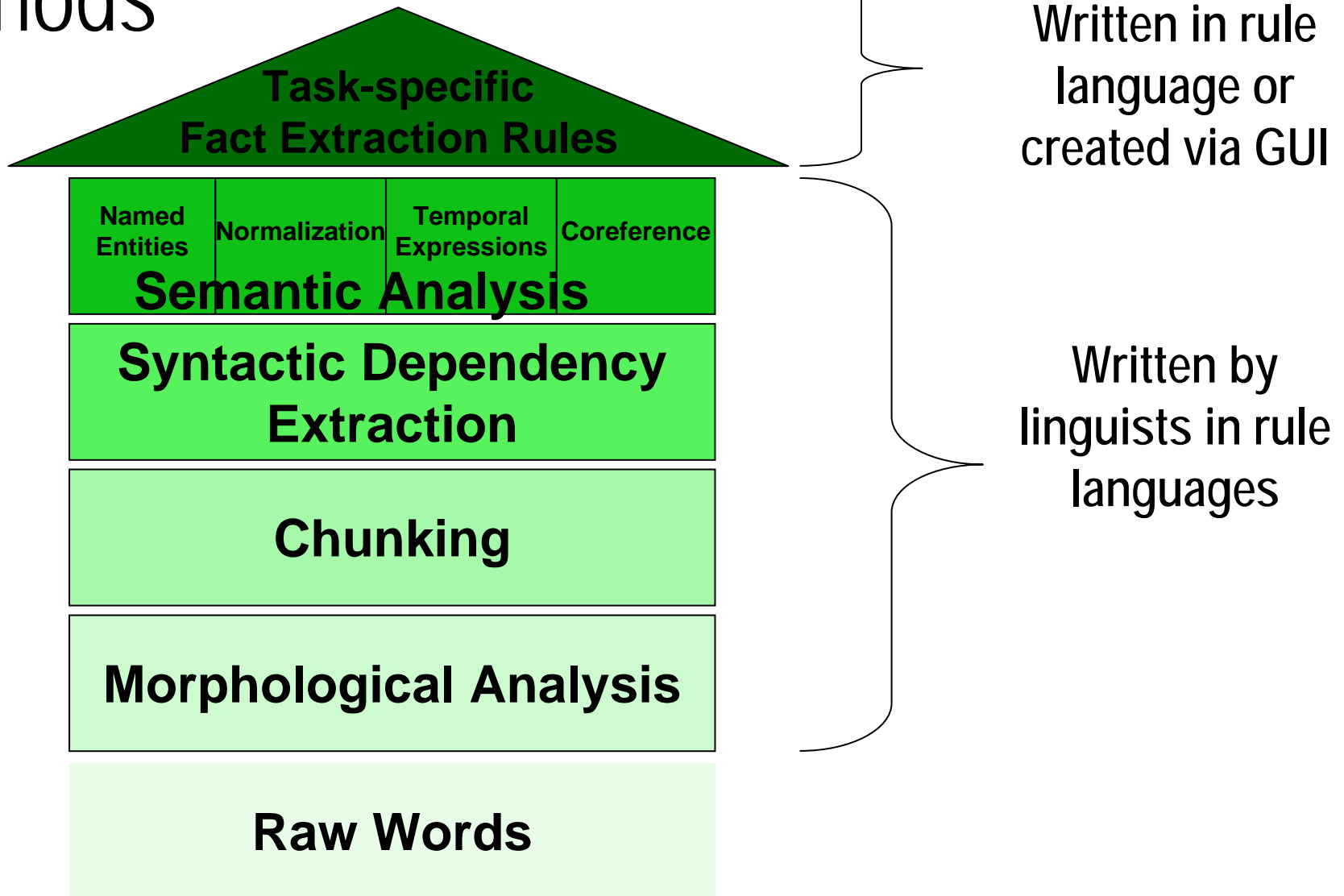
Bush<sub>7</sub> said Wednesday that the United States<sub>5</sub> and its<sub>6</sub> European allies would seek UN<sub>6</sub> **Security Council**<sub>95</sub> action against Iran<sub>34</sub> if **Tehran**<sub>97</sub> rejected incentives to limit its<sub>97</sub> nuclear programs. *"The understanding is, we<sub>7</sub> go to the Security Council<sub>95</sub> if they<sub>97</sub> reject the offer. And I<sub>7</sub> hope they<sub>97</sub> don't. I<sub>7</sub> hope they<sub>97</sub> realize the world is clear about making sure that they<sub>97</sub> don't end up with a nuclear weapon," the US<sub>5</sub> president<sub>7</sub> said.*

**Quotation: President George W. Bush**<sub>7</sub>

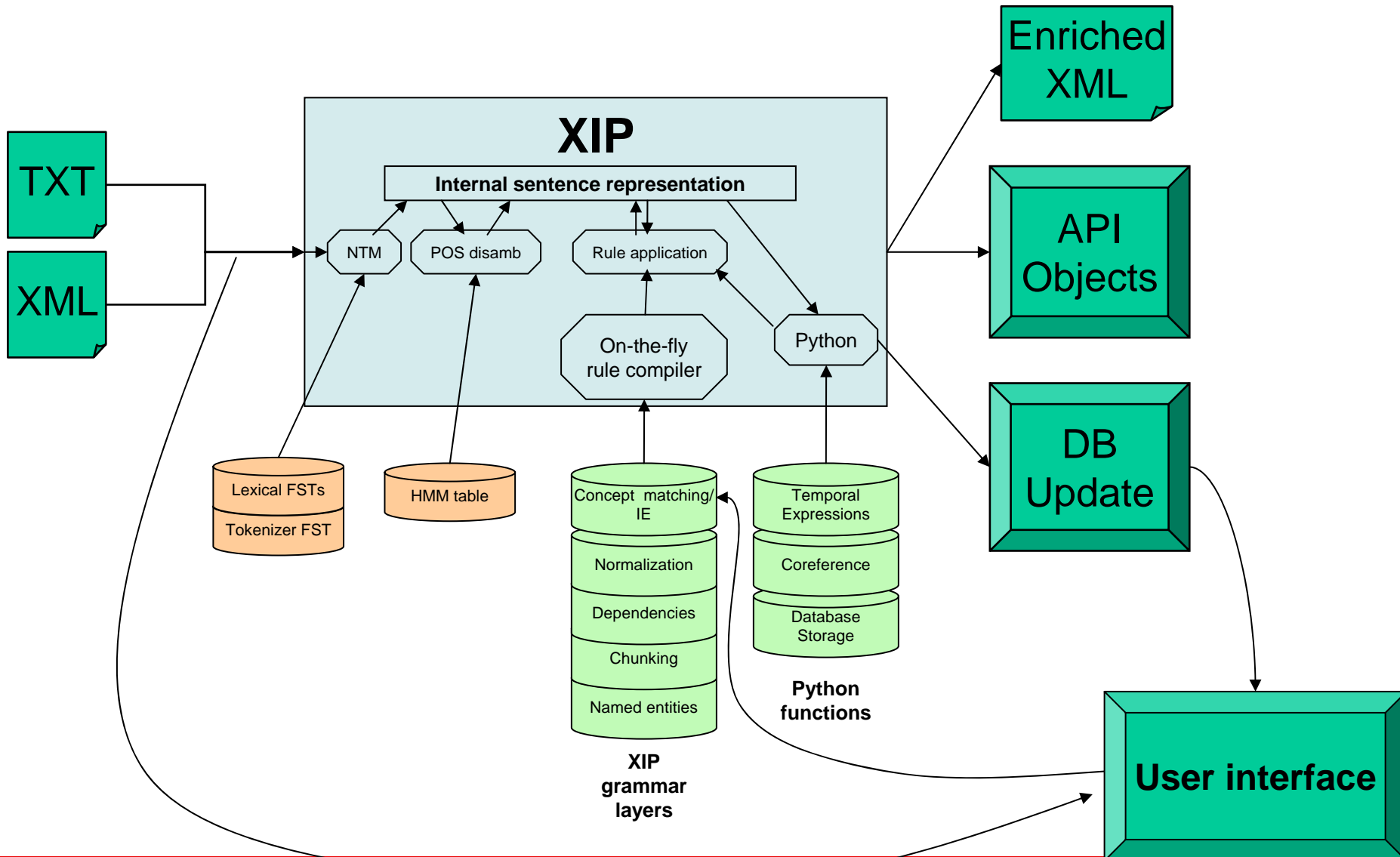
# Hierarchy of Information Discovery **XEROX**

*Research Centre Europe*

## Methods



# Architecture Overview



# Languages Available

Morphological analysis: Arabic, Chinese, Czech, Dutch, English, French, German, Hungarian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Turkish

Grammars:

- Currently available: English, French
- Coming soon: German, Italian, Portuguese, Spanish