XEROX
Research Centre Europe

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

XEROX
Research Centre Europe

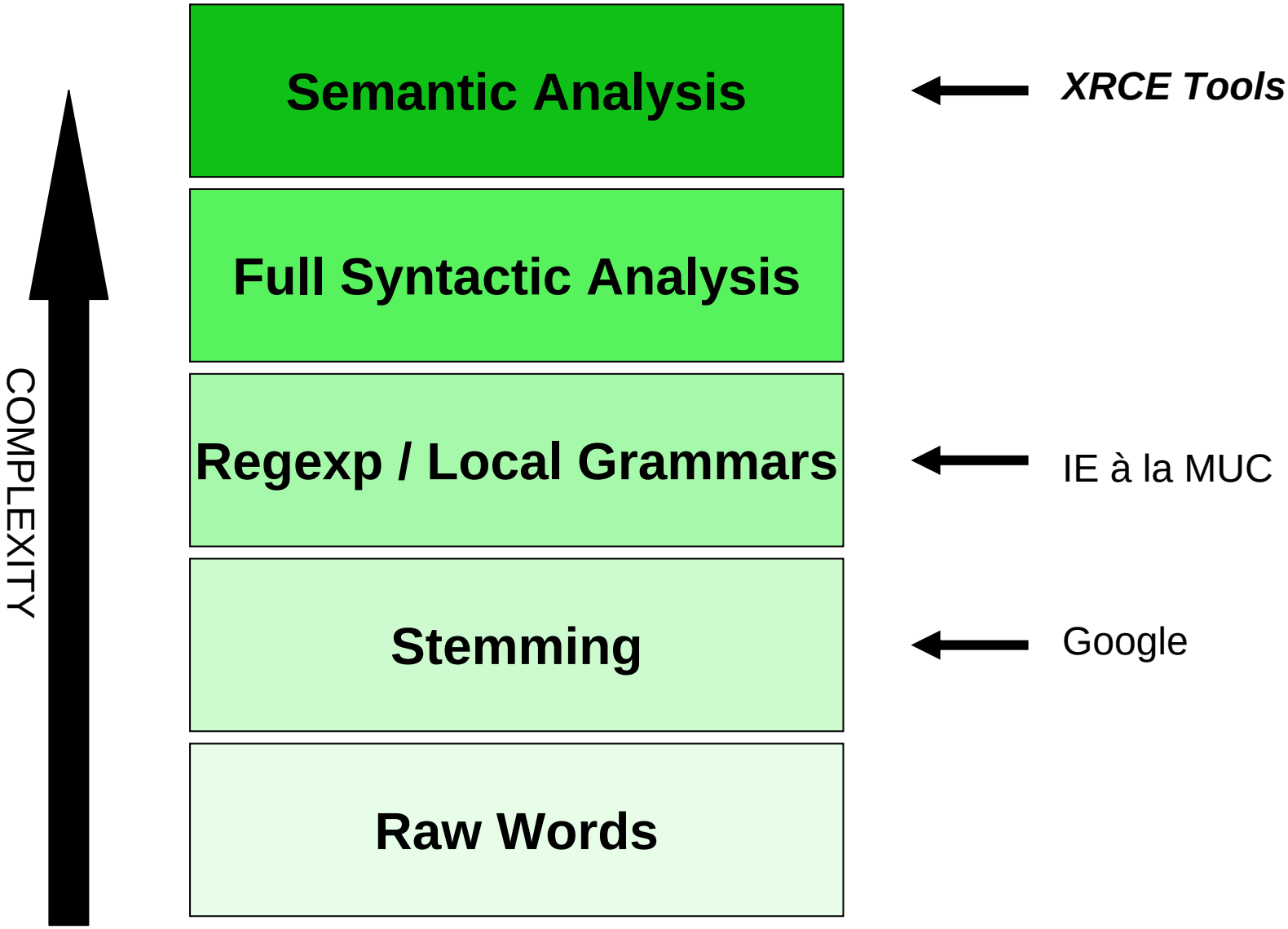
Information Retrieval: finding documents on a given topic, or documents that might answer a given question

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
Research Centre Europe



Xerox Incremental Parser (XIP)

XEROX
Research Centre Europe

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

XEROX
Research Centre Europe

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

- Tokenization & morphological analysis (FST)

Incremental Syntactic Analysis

XEROX
Research Centre Europe

Hervé	*Gallaire*	*joined*	*Xerox*	*in*	*1992*
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

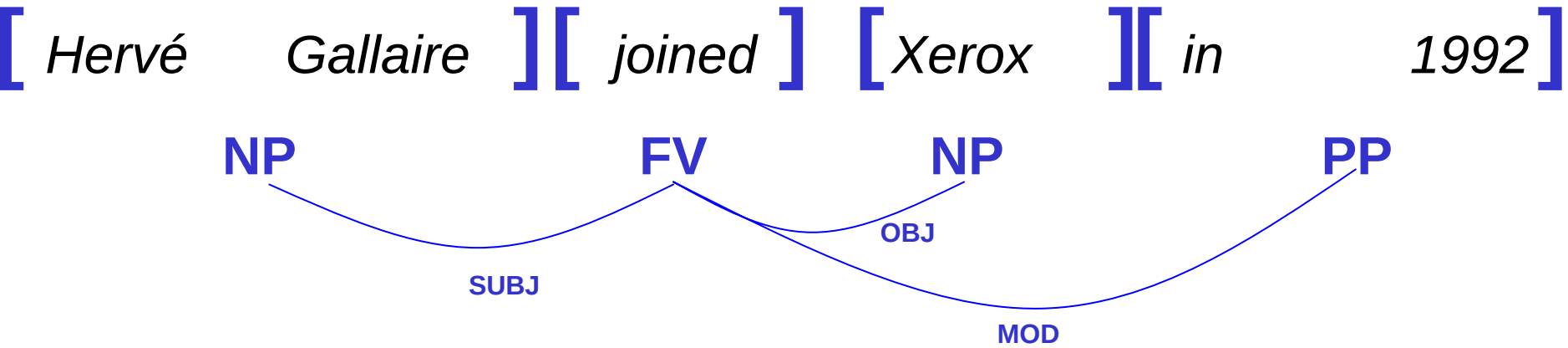
XEROX
Research Centre Europe

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

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

Incremental Syntactic Analysis

XEROX
Research Centre Europe



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

Semantic Analysis: Named Entities

XEROX Research Centre Europe

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 Marymount 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

XEROX
Research Centre Europe

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*	PURPOSE
*A train **for** London*	DESTINATION
*He has been waiting **for** hours*	DURATION

Semantic Analysis: Temporal Expressions

XEROX
Research Centre Europe

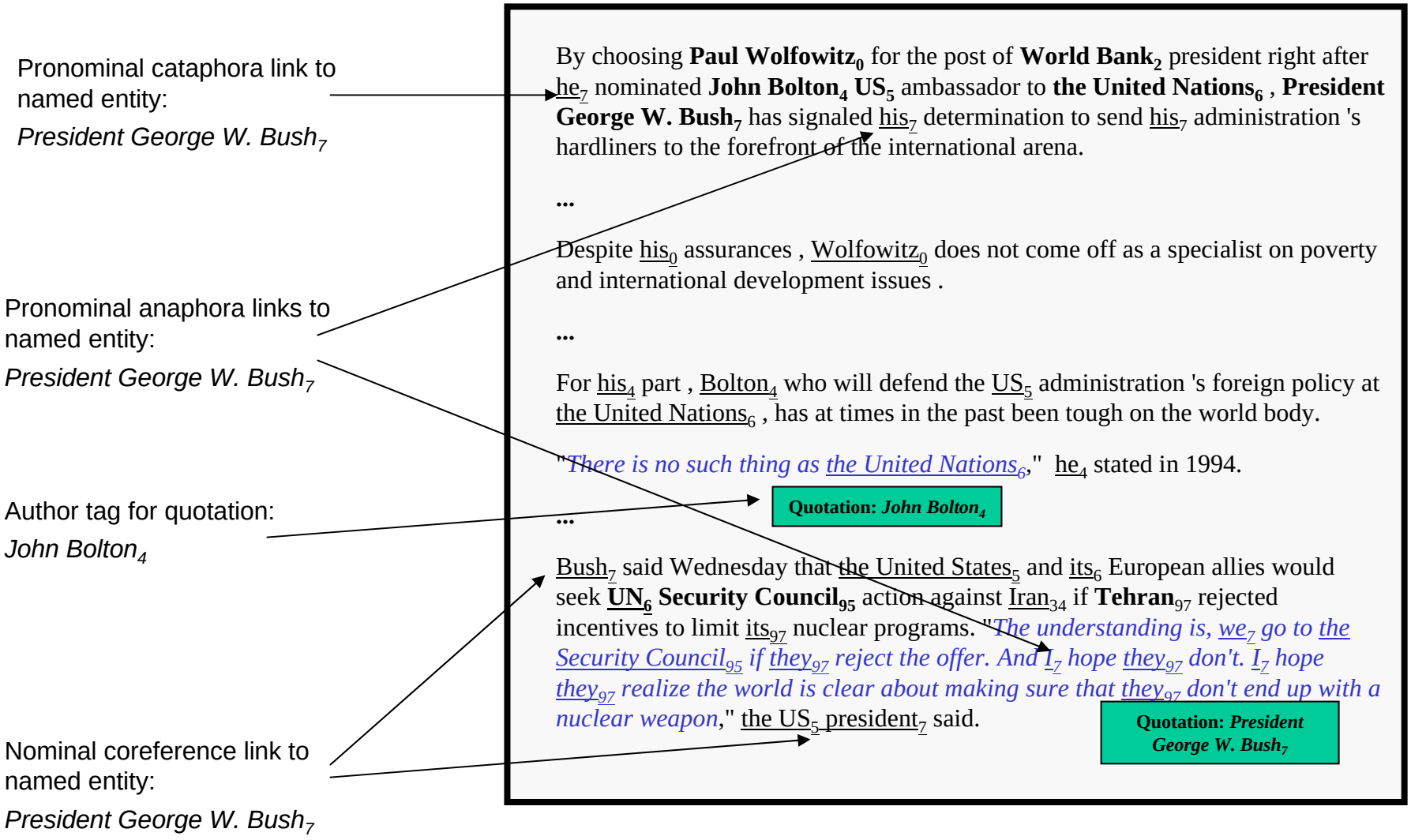
Hervé J. Gallaire is president of the Xerox Innovation Group and Xerox's chief technology officer. He was appointed to this position in October 2001 and was named a corporate senior vice president in December 1999. … He joined Xerox in 1992 and managed the Xerox Research Centre Europe in Grenoble, France,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

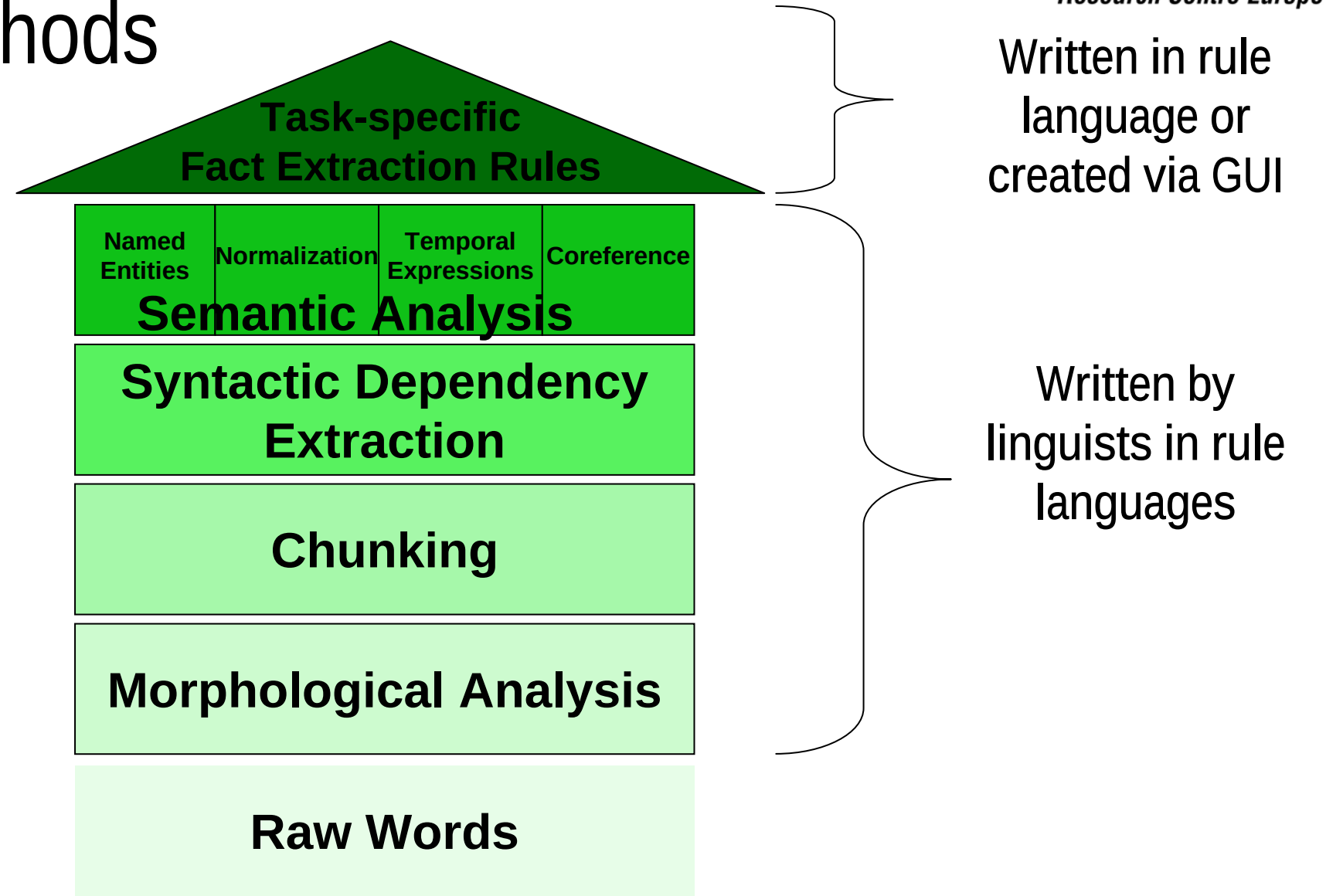
Semantic Analysis: Coreference

XEROX Research Centre Europe



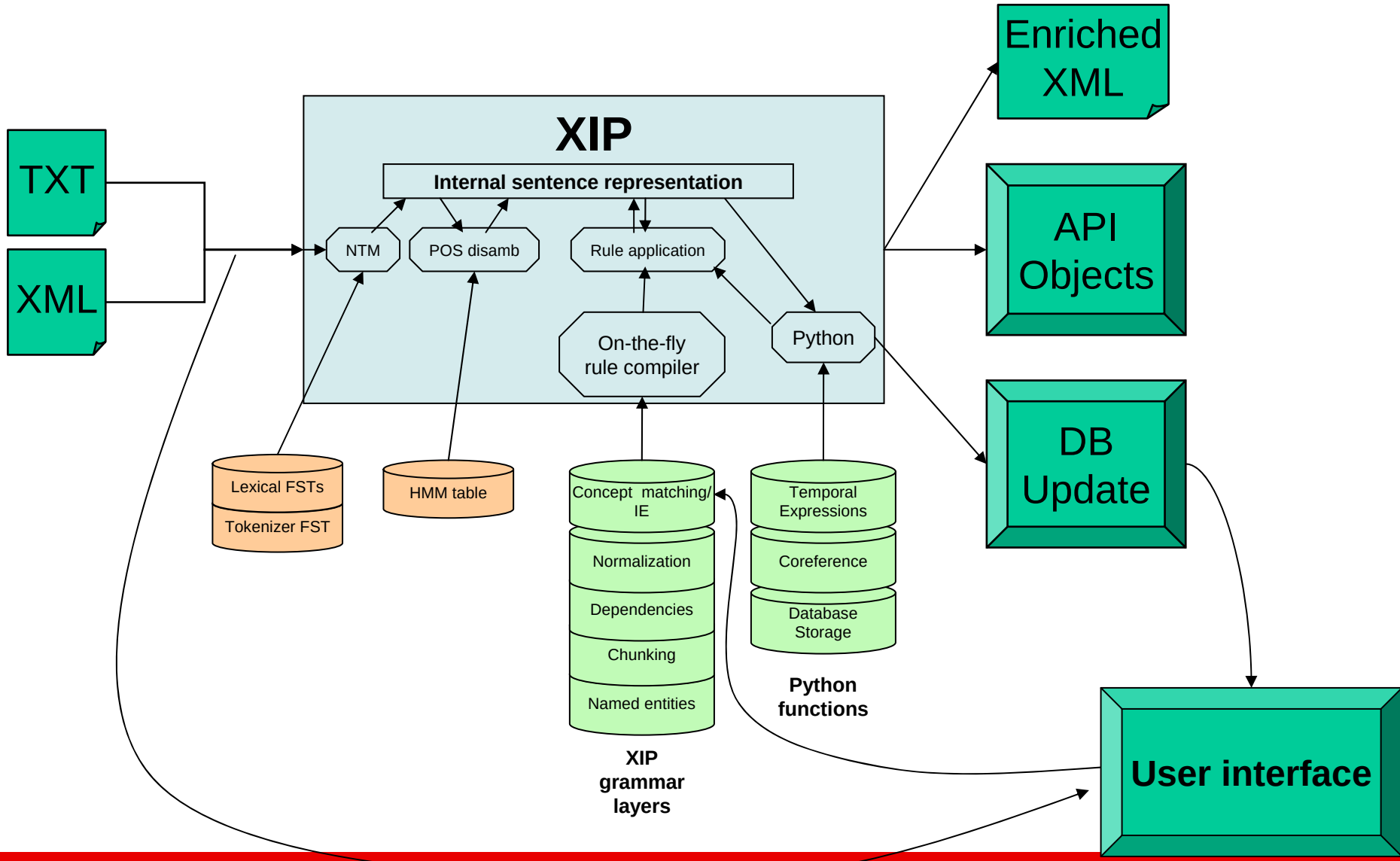
Hierarchy of Information Discovery Methods

XEROX Research Centre Europe



Architecture Overview

XEROX
Research Centre Europe



Languages Available

XEROX
Research Centre Europe

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