

ALVIS

A Robust Linguistic Infrastructure for Efficient Semantic Content Analysis

Thierry Poibeau

LIPN, CNRS et Université Paris 13

Joint work with A. Nazarenko, T. Hamon,
S. Aubin, J. Derivière, D. Weissenbacher





Overview

- ALVIS: STREP FP6 project
- Coordination: Wray Buntine (Helsinki University of Technology)
- Goal: Building an infrastructure for open source search engines, using a peer-to-peer and subject-specific technology
 - P2P Large-scale Information Retrieval
 - Subject specific Information Retrieval



ALVIS consortium

Participant name	Country Short name	Technical Lead
Helsinki University of Technology, Helsinki Institute for Information Technology HIIT	Finland / HUT	Wray BUNTINE
Institut National de la Recherche Agronomique, Unité Mathématique, Informatique et Génome	France / INRA	Claire NEDELLEC
Ecole Polytechnique Fédérale de Lausanne, Distributed Information Systems Lab	Switzerland / EPFL	Karl ABERER
Lund University, Department of Information Technology	Sweden / ULUND	Anders ARDÖ
Technical University of Denmark, Center of Knowledge Technology	Denmark / DTU	Gert SCHMELTZ PEDERSEN
Index Data Aps	Denmark / INDEX DATA	Marc CROMME
Exalead SA	France / EXALEAD	Francois LAGUNAS
Université Paris-Nord, Laboratoire d'Informatique	France / PARIS 13	Adeline NAZARENKO
ALMA Bioinformatics, S.L.	Spain / AB	Christian BLASCHKE
Jozef Stefan Institute, Department of Intelligent Systems	Slovenia / JSI	Marko GROBELNIK
Tsinghua University, Department of Computer Science and Technology	China / TU	Lizhu ZHOU



Rationale & Objectives

- Development path based on open source
- Collaboration between Information extraction and data mining
- Target specific user categories, don't compete with the majors
- Enable different user experiences with simple semantic capability
- Empower European-centric search initiatives



A P2P approach

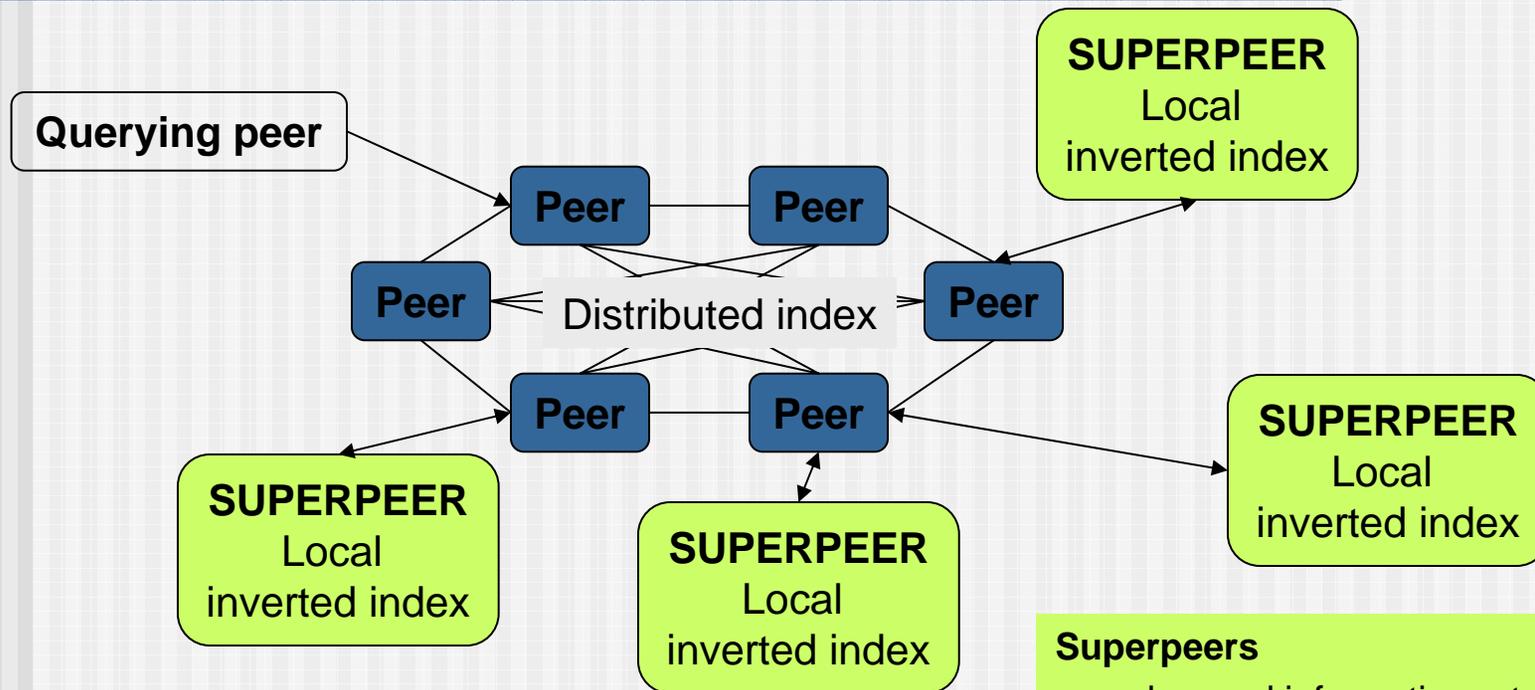
Rationale

- scalability
- self-organisation
- fault-tolerance

Architecture with two types of nodes

- *Peers* building the distributed P2P overlay network
- *Superpeers*, stand-alone components hosting document collections

Architecture Overview



Peers

- maintenance of an inverted index of superpeers' document collections
- efficient querying of its distributed index

Superpeers

- advanced information retrieval services
- sophisticated processing of document collections to build semantically rich indexes enhanced by various ranking strategies
- complex structured queries

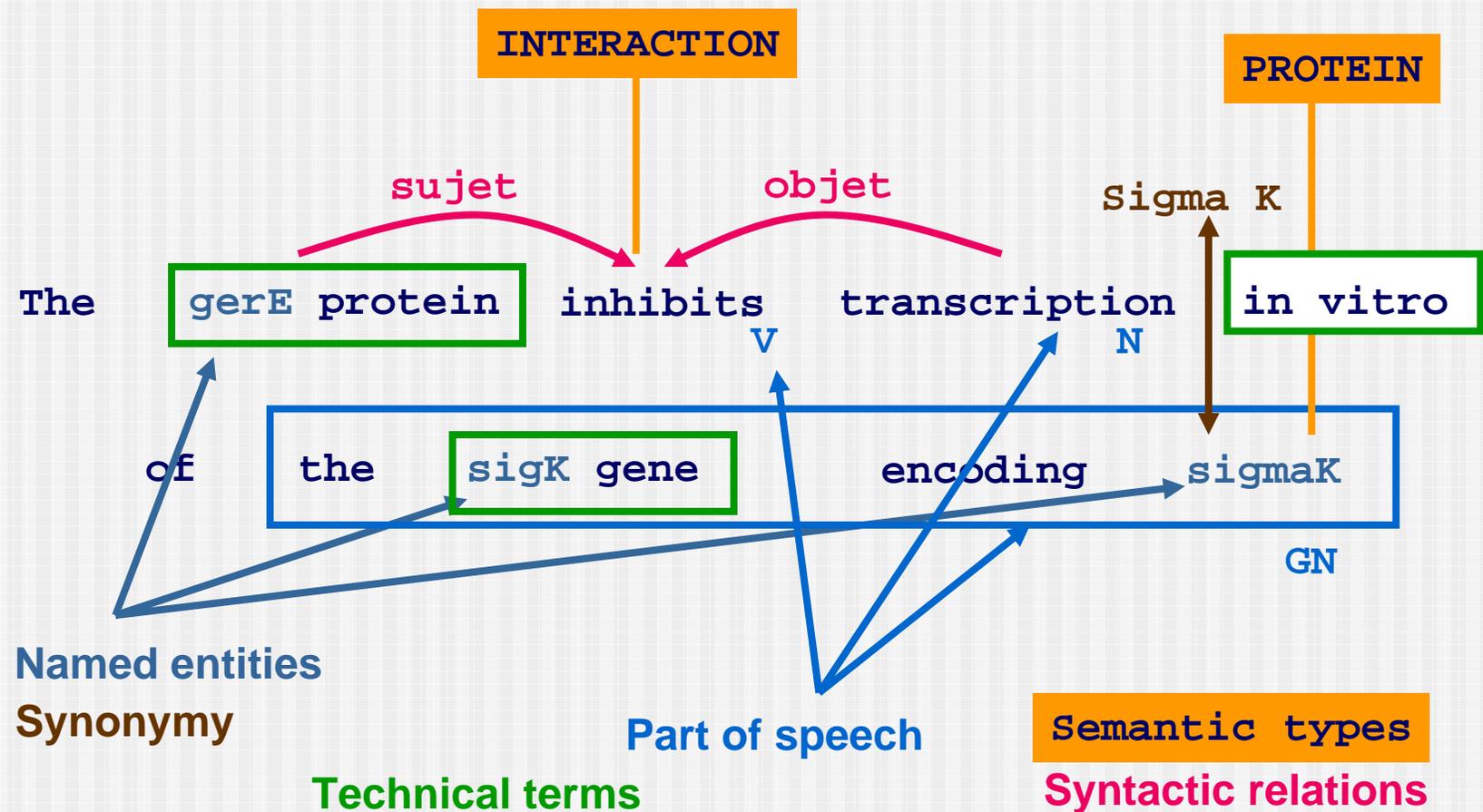


Text analysis

- University Paris 13 is mainly involved in natural language processing (NLP) for IR
- Semantic annotation
 - Technical term analysis
 - Named entity recognition
- Extraction of relevant information (IE)
 - Acquisition of structured information from texts

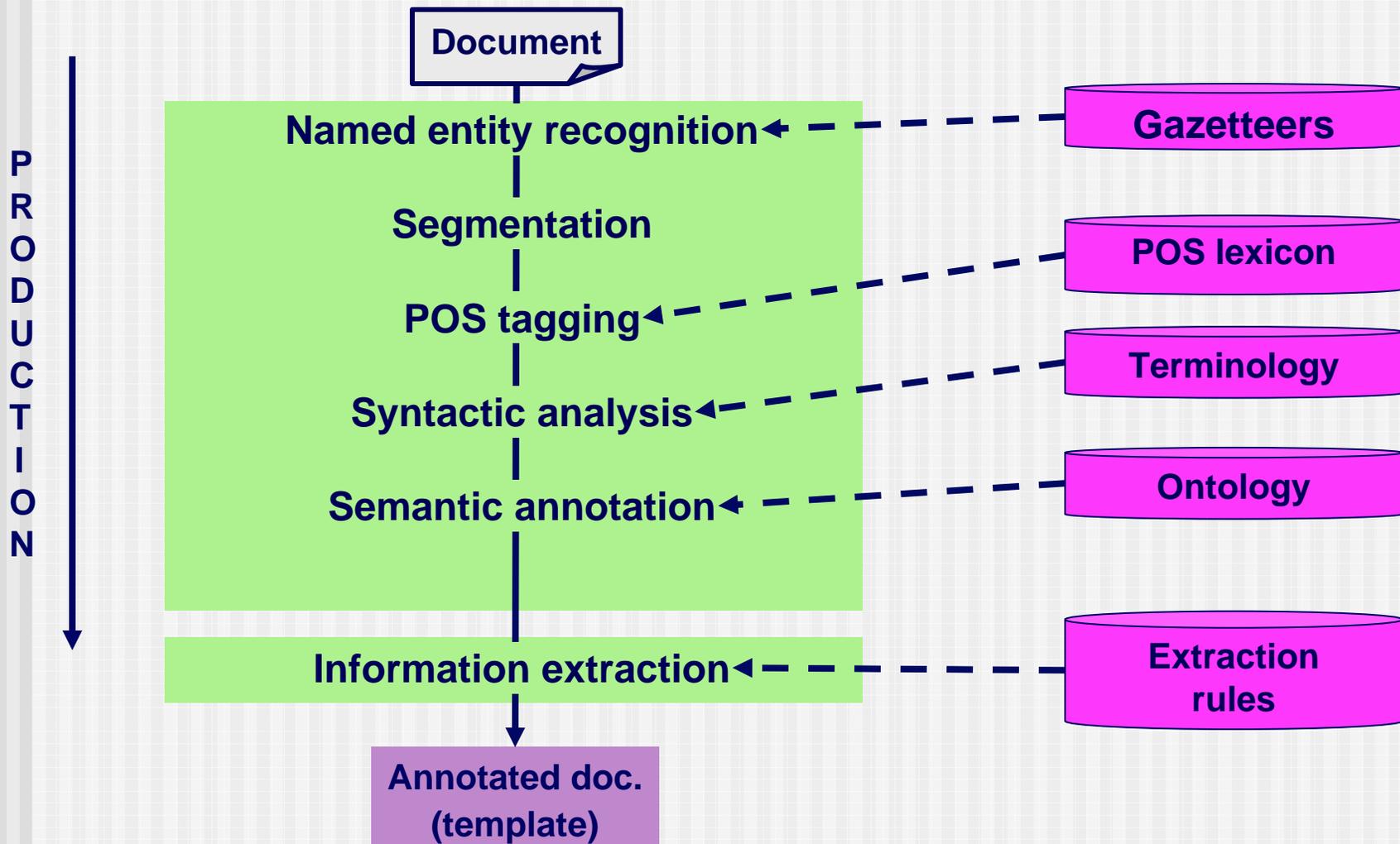


Annotations layers





NLP Architecture





Recycling existing NLP tools

POS tagging	Brill, TreeTagger	« inhibits » is a verb
Syntactic analysers	IFSP, Link Parser	« gerE protein » is the subject of « inhibits »
Term extractors	ACABIT, SYNTEX	« in vitro » is a technical expressions

NLP tools achieve poor results when applied on specialized corpora

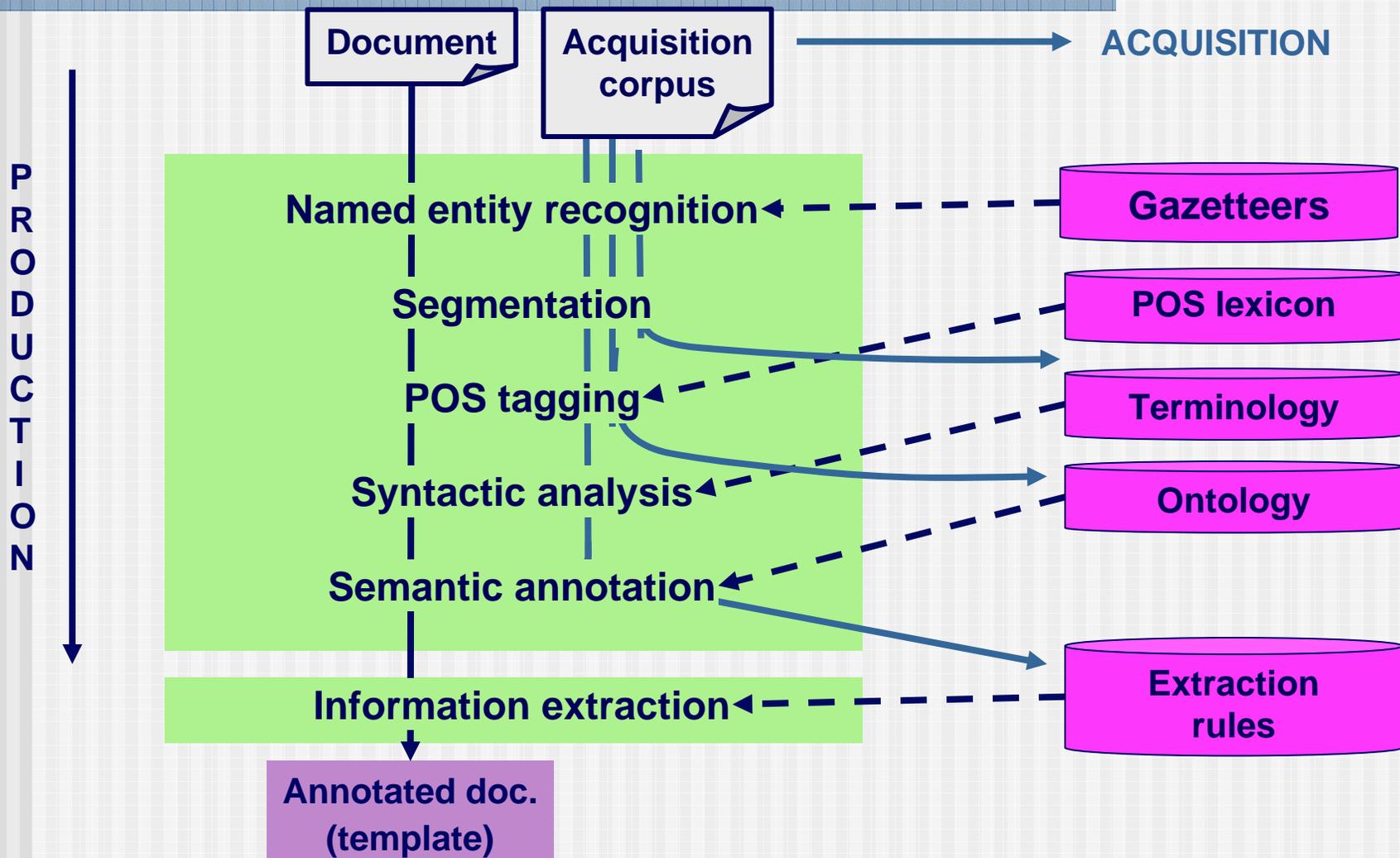
- Subject-verb relation: precision < .7
- Coordination: precision < .4



Tool and resource tuning

- Objective: a quick and efficient adaptation of resources and tools for a new domain
- Strategy
 - Acquire domain knowledge from representative corpora
 - Enhance tool performance with this specific knowledge (terminology, specific grammars...)
 - Use all the information available from previous analysis steps

Mixing acquisition with production



Conclusion and perspectives



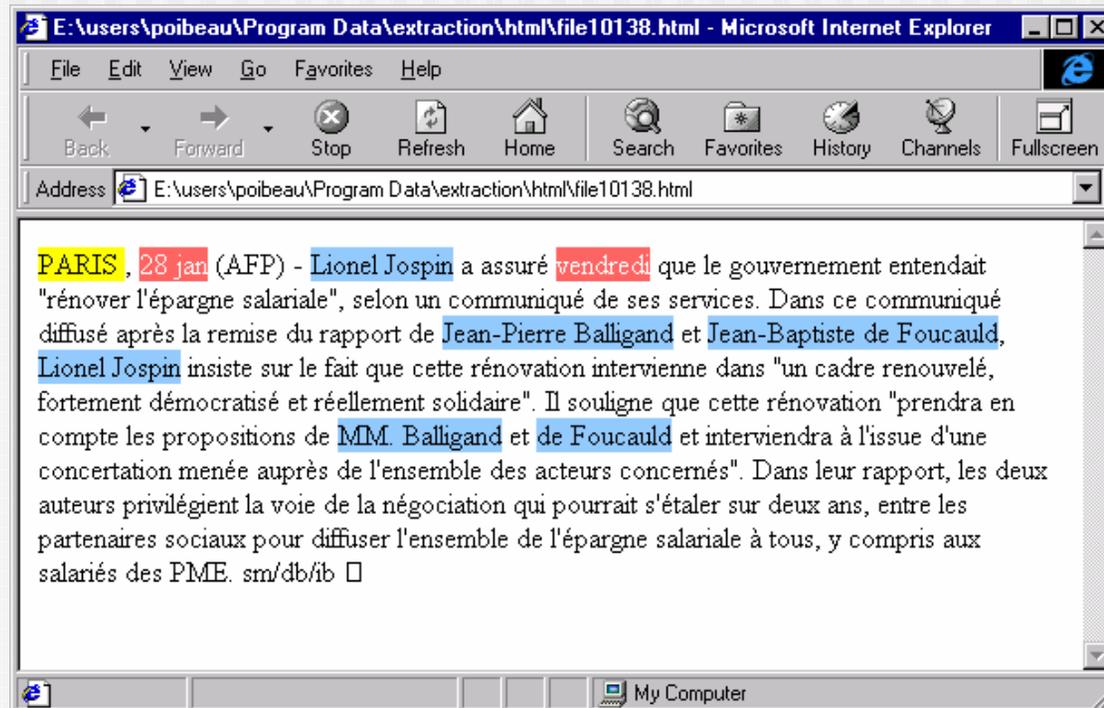


Overview...

- An annotation platform integrating different annotation layers
 - Annotation standards (UTF-8, ISO TC37/SC4)
 - Open source software
 - Multilingualism (mainly English and French, but Chinese and Slovenian ongoing)
 - Scalability (Towards Gb)
- Tool and resource tuning
 - Named entity, term analysis...
 - Syntactic analysis for accurate IE performances



Semantic annotation



Multilingualism



The image displays four overlapping browser windows, each showing a different language's text with named entities highlighted in blue. The windows are titled "Named entities - Microsoft Internet Explorer".

- Top-left window (Arabic):** Shows text about the 1922 election in Egypt. Highlighted entities include "مبارك" (Nasser) and "الأمم المتحدة" (United Nations).
- Top-right window (Russian):** Shows text about the UN Security Council. Highlighted entities include "США" (USA), "президент" (president), "Ирак" (Iraq), "Шведский дипломат" (Swedish diplomat), "Hans Blix", "Афганистане" (Afghanistan), and "ПОЛЬ-МАРИ ДЕ ЛА ГОРС" (Pol-Marie de La Gorce).
- Bottom-left window (Russian):** Shows text about the UN Security Council. Highlighted entities include "США" (USA), "Hans Blix", "Атомовой (МАЕА)", "Bagdad", "Blix", and "Baradei".
- Bottom-right window (Japanese):** Shows text about the 2004 Madrid train bombing. Highlighted entities include "スペイン" (Spain), "イグナシオ・ラモネ (Ignacio Ramonet)", "ル・モンド・ディプロマティーク編集総長" (L'Espresso editor-in-chief), "佐藤健彦" (Shimoda Kenji), "3月11日" (March 11), "イラク戦争" (Iraq War), "アルカイダ" (Al-Qaeda), "フッシュオ大統領" (Bush), "ワウシアラ" (Wushera), "モロッコ" (Morocco), "トルコ" (Turkey), and "EU" (EU).

Multilingual named entity recognition



Information extraction

afp_0001.txt - Bloc-notes

Fichier Edition Format Affichage ?

PARIS, 5 oct (AFRP) - Le Premier ministre français Lionel Jospin a affirmé jeudi que "M. I éviter un peuple".

Vojislav l'Opposi (DOS), s libérée" manifest à Belgra fier d'a Yougosla certaine: dans les terrasse

Date	28-ja-00
Location	PARIS
Personality	Lionel Jospin
Type	Déclaration
Source	AFP (wire)
Topic	- Lionel Josp gouvernement l'épargne sal communiqué de

e:\users\poibeau\program data\extraction\txt\file10138.txt

Date	31-ju-00
Location	AUXERRE
Sport	football
Type	Résultat sportif
Source	AFP (wire)
Topic	- "Maintenant, il me faudra réapprendre à perdre". Lors de sa nomination pour succéder à Guy Roux comme entraîneur de l'AJ Auxerre, Daniel Rolland avait eu ce bon mot. Samedi soir à domicile, pour son

e:\users\poibeau\program data\extraction\txt\file10139.txt

Date	
Location	
Sport	
Type	
Source	
Topic	