Oberon for Natural Language Processing

Eric Wehrli & Luka Nerima
LATL-Dept. of Linguistics
University of Geneva
Eric.Wehrli@lettres.unige.ch
http://www.latl.unige.ch

Oberon day @ CERN March 10, 2004

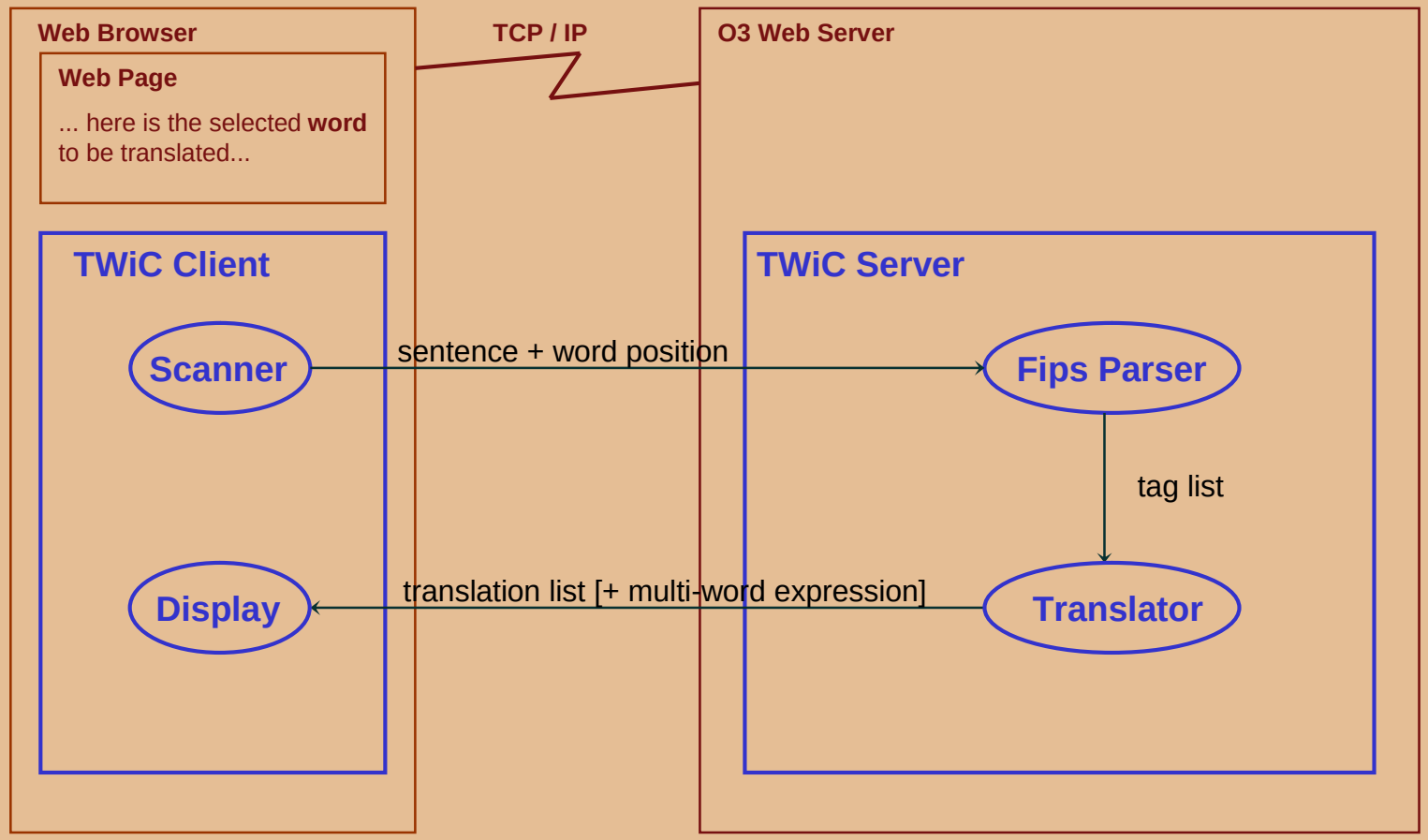
A concrete example : TWiC 1-The problem

- Provide terminological assistance to readers of on-line documents in foreign languages.
- Neither on-line dictionaries nor machine translation constitute adequate solutions:
  - Dictionaries tend to be « noisy » (ignoring contextual information, they return irrelevant information)
  - Machine translation is still too unreliable

2-Proposed solution

- TWiC (Translation of words in context) is a bilingual (English-French) terminological assistant for on-line documents.
  - Given a selected word, TWiC will display possible translations compatible with the linguistic context (syntax and very partially semantics)
  - For instance, given the word « *gave* » in « *he gave it up* », TWiC returns « *abandonner, renoncer à* » and not the dozens of possible translations of the verb « to give »

TWiC Architecture



TWiC POS tag list

They foiled an attempt…

Source word	POS tag	Position	Lexeme number	Expression number
they	PRO-PER-3-PLU	0	111000011	
foiled	VER-PAS-3-PLU	5	111016454	141000136
an	DET-SIN	12	111050002	
attempt	NOU-SIN	15	111005034	- 141000136

Advantages

- Better identification of selected item (less noise)
  - Ils ont passé tout l'été (summer vs been)
  - They all rose
- Identification of multiword expressions
  - They didn't get along well.
  - The record she has broken was 10 years old.

$[_{DP}$ the $[_{NP}$ record$_i$ $[_{CP}$ $[_{DP}$ e $_i$ $]$ $[_{TP}$ $[_{DP}$ she $]$ has $[_{VP}$ broken $[_{DP}$ e $_i$ $]]]]]]$

  - They saw a school of little fishes.
  - He foiled an attempt.

Some figures

- Size of lexical DB
  - French & English monolingual dictionaries :
    - ~50k lexemes + ~2500 expressions
    - >200k morphological forms (>100 for English)
  - Bilingual (English-French): ~50k entries
- Proc. speed : ~150 words/sec
- Size of application
  - Client module : ~1MB
  - Server module : ~2,5MB
  - ISAM datafiles : ~40MB
- Fips source code (generic)
  - 35 modules, ~37'500 lines of code
- Source code (language-specific)
  - 2 modules, ~7'000 lines of code (per language)

Why Oberon ? Why BlackBox ? (1/2)

- **Automatic garbage collection**

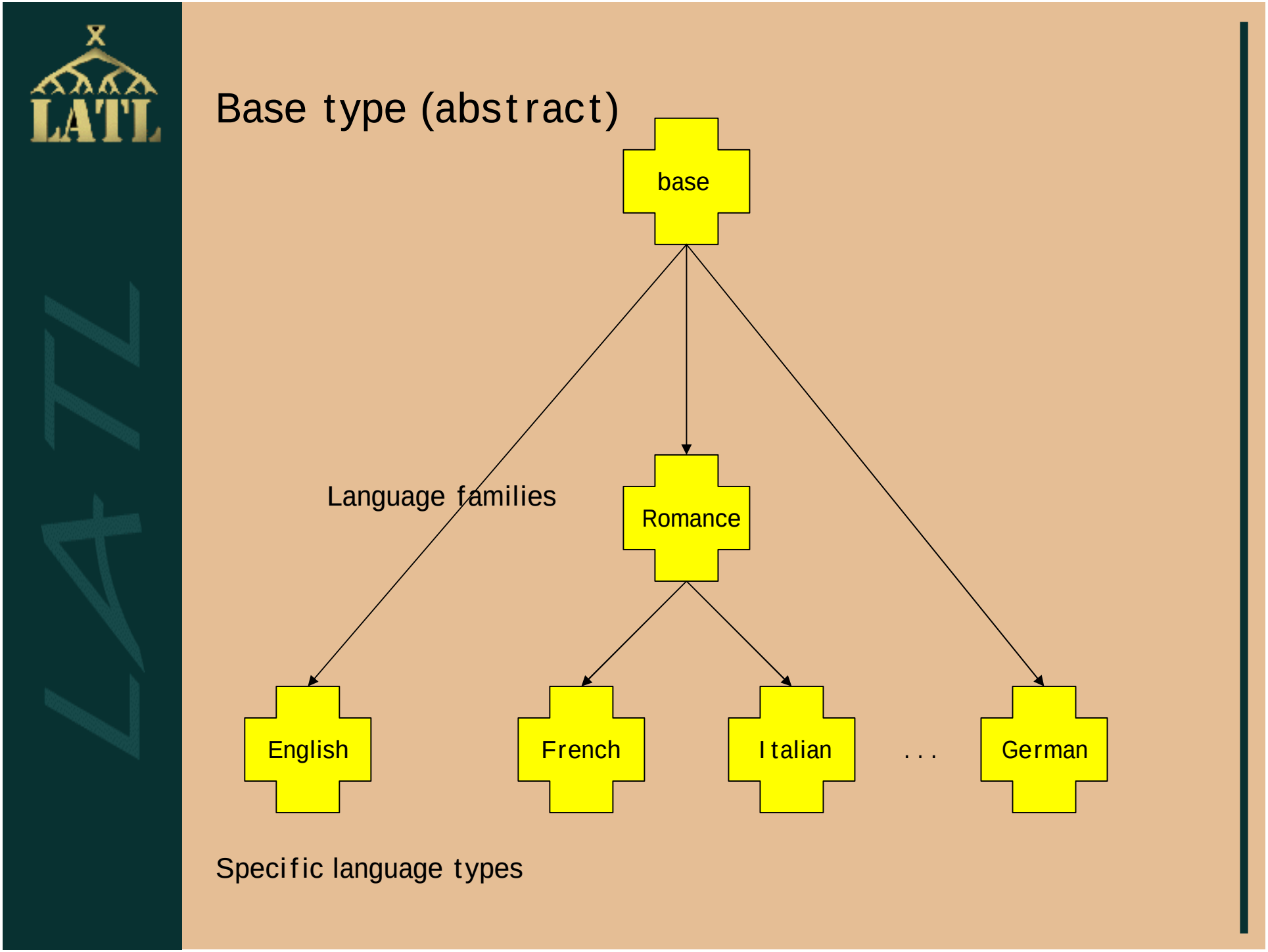
  NLP is hugely non-deterministic (combinatorics of syntactic ambiguities such as prepositional phrase attachments corresponds to the Catalan number sequence)

- **Fast code** (vs Prolog or Lisp)

  Given the high-level of non-determinism of NLP applications, extremely fast code is necessary to achieve real time responses

- **Object-oriented language**

  Object design appears to be a good/interesting way to model language variation


Base type (abstract)
base
Language families
Romance
English
French
Italian
. . .
German
Specific language types

Why Oberon ? Why BlackBox ? (2/2)

- Environment is fully unicode and well-integrated in the Windows system we have done some morphological work on Greek, Hungarian, Russian, and would like to consider Asian and Semitic languages
- Easy to develop distributable exe or dll components
- Hypertext facilities and more generally the richness of the MVC design
- Top-level assistance and support from OM