THE SIXTH FRAMEWORK PROGRAMME
The sixth framework programme covers Community
activities in the field of research, technological
development and demonstration (RTD)
for the period 2002 to 2006

D5.3 CLIR/COMPREHENSION AID-SOFTWARE FOR FIRST USER TRIALS

-

Version 1.0
4/Jul/2008

Executive Summary

VERSION	DATE	AUTHORS
01	04/07/2008	Kimmo Valtonen, Vladimir Poroshin

TITLE	D5.3 CLIR/comprehension aid-software for first user trials
STATE	Draft
CONFIDENTIALITY	PU
AUTHOR(S)	Kimmo Valtonen, Vladimir Poroshin
CONFIDENTIALITY	
PARTICIPANT PARTNERS	University of Helsinki, Amebis, University of Southampton, JSI, UCL, XRCE
WORKPACKAGE	WP5
ABSTRACT	This document provides documentation of the prototype
KEYWORDS	
REFERENCES	
COMMENTS	
REVIEWER	Jean-Michel RENDERS, Nicola Cancedda, Blaz Fortuna

Contents

Chapter 1

Description of the system

The prototype is designed to be modular, with each module accessible via TCP or HTTP for maximum ease of access. This also frees the system of any need for the modules to reside at the same physical location. It further enables parallel development of the modules at different sites, as long as all implementations obey the protocol. Due to readability and availability of standardized processing tools, XML has been chosen as the format for messages (requests and replies) passed between the modules. The choice of programming language and operating system to run on is also left entirely up to the site implementing a particular module.

The system consists of these parts:

- A query translation/mapping module.
- An HTML translation module that attempts to retain the original formatting and linkage while translating the textual content.
- The translation module.
- The GUI (Graphical User Interface).

See Fig. 1.1 for a schematic representation of the architecture

The user accesses the Graphical User Interface via HTTP (using a browser). The interface then accepts the user's queries and issues first a Search request over TCP to the central component to do cross-lingual information retrieval. The resulting hits (documents in language B matching the query formulated in language A) can then be translated at will by issuing a Translate request to the central component.

To carry out these requests, the central component accesses the other three modules listed above, as shown.

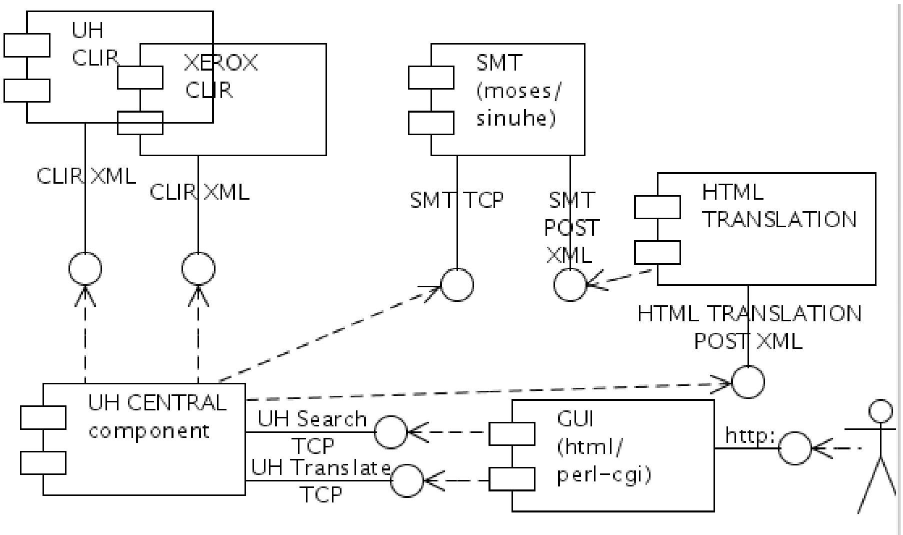


Figure 1.1: Overall view of the system architecture.

Chapter 2

Using the prototype

The prototype can be found at the URI

```
http://cosco-demo.hiit.fi/smart/
```

Queries are currently only possible in Slovenian as the current version reflects the Use Case where the user issues queries in Slovenian and receives results from the French Wikipedia, machine translated to Slovenian. We also offer the possibility to translate from French to English for those with no competence in Slovenian to allow assessing the performance in terms of quality.

2.1 General instructions

Let us try for example the query

```
evropski parlament
```

("European parliament" in Slovenian). What you will see for each resulting hit is a link to the translated version above a link to the original French Wikipedia article. The Wikipedia link is to the actual on-line version as it existed in the recent past. See Fig. 2.1 and Fig. 2.2.

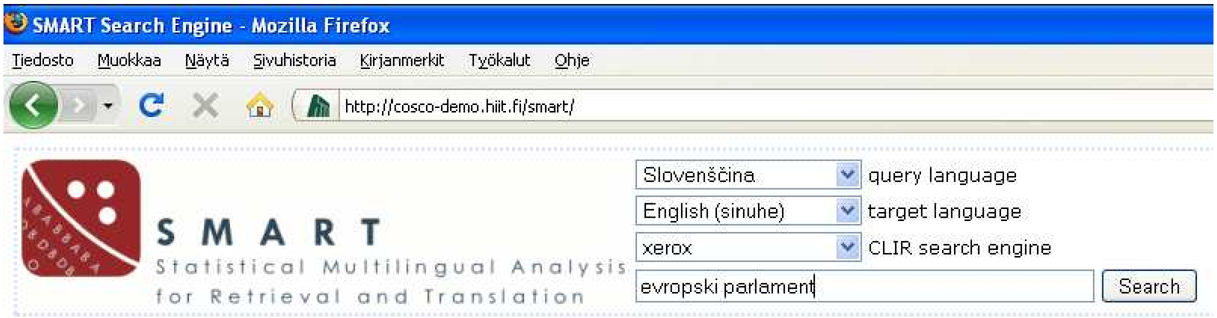
The title of the translation page is already translated (on-line). Snippets are also provided. If you click on the translated title, the HTML translation module will then call the translation server and the article will be translated on-line, retaining the original formatting.

The original article's current on-line version (if one exists) can be found by clicking on the title in French, see Fig. 2.3.

Note that the on-line version is a living thing and might differ from the one used for translation by our system. The translation obtained will look like in Fig. 2.4.

2.2 Choosing the target

By default, the system will currently translate using a model built using the Moses software, which is not a product of the SMART project. However, the user can also choose to use some other translation method or target language in the "target language" box. The set of available options are English or Slovenian as the target language. As the decoder the user can at the same time pick either Moses or a SMART method (Sinuhe or Southampton (to appear soon)).


SMART Search Engine - Mozilla Firefox
Tiedosto
Muokkaa
Näytä
Sivuhistoria
Kirjanmerkit
Työkalut
Ohje
http://cosco-demo.hiit.fi/smart/
S M A R T
Statistical Multilingual Analysis
for Retrieval and Translation
Search

Slovenščina query language

English (sinuhe) target language

xerox CLIR search engine

evropski parlament

Figure 2.1: Entering the query in Slovenian.

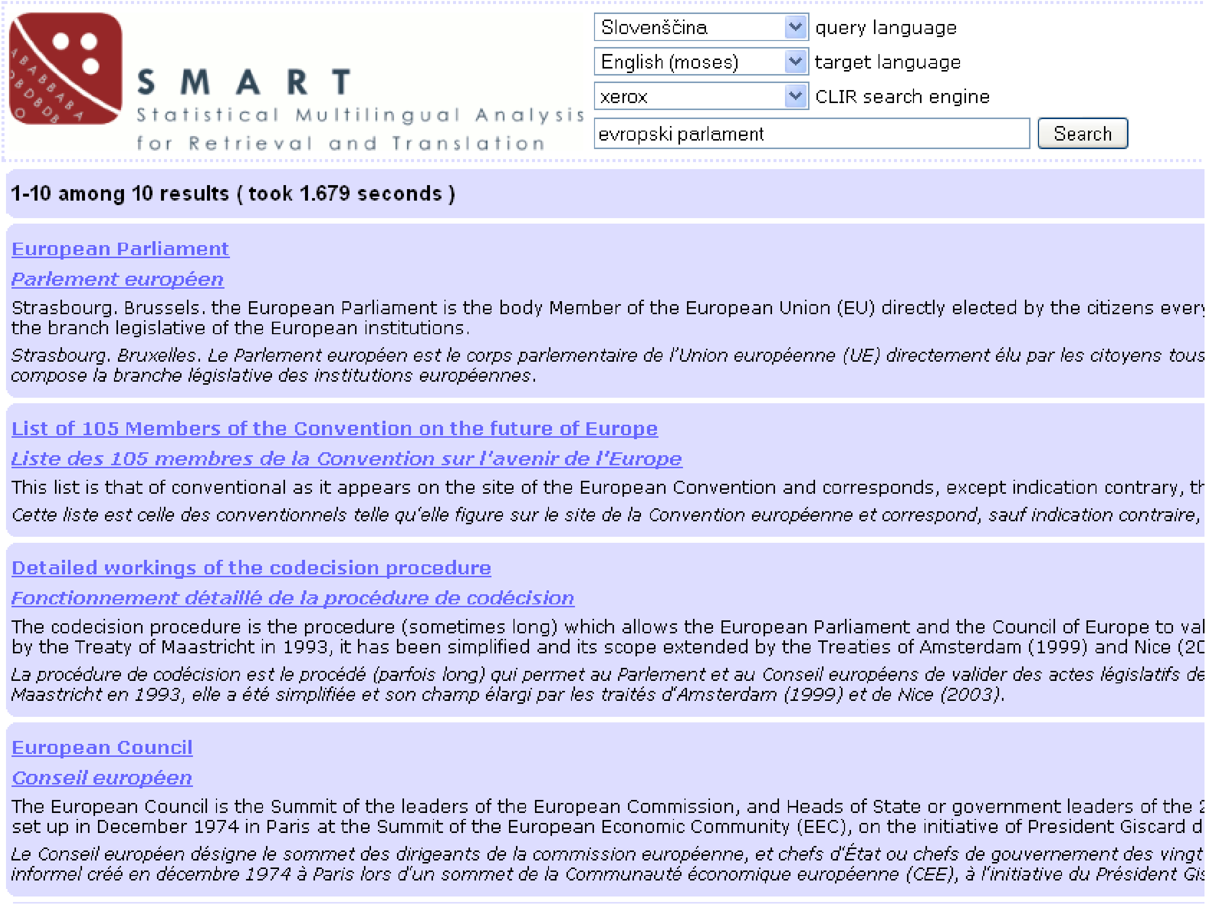
2.3 Choosing the CLIR method

The CLIR engine used can also be chosen from among alternatives. Currently the default is Xerox's query mapping software, but a method developed at Univ. of Helsinki can also be chosen.

2.4 Arbitrary queries

To test the system further using other query words, you can use Amebis' online translation service from English to Slovenian:

```
http://presis.amebis.si/prevajanje/
```


SMART
Statistical Multilingual Analysis
for Retrieval and Translation
Search

Slovenščina query language

English (moses) target language

xerox CLIR search engine

evropski parlament

1-10 among 10 results (took 1.679 seconds)

European Parliament

Parlement européen

Strasbourg. Brussels. the European Parliament is the body Member of the European Union (EU) directly elected by the citizens every
the branch legislative of the European institutions.

*Strasbourg. Bruxelles. Le Parlement européen est le corps parlementaire de l'Union européenne (UE) directement élu par les citoyens tous
compose la branche législative des institutions européennes.*

List of 105 Members of the Convention on the future of Europe

Liste des 105 membres de la Convention sur l'avenir de l'Europe

This list is that of conventional as it appears on the site of the European Convention and corresponds, except indication contrary, th

Cette liste est celle des conventionnels telle qu'elle figure sur le site de la Convention européenne et correspond, sauf indication contraire,

Detailed workings of the codecision procedure

Fonctionnement détaillé de la procédure de codécision

The codecision procedure is the procedure (sometimes long) which allows the European Parliament and the Council of Europe to val
by the Treaty of Maastricht in 1993, it has been simplified and its scope extended by the Treaties of Amsterdam (1999) and Nice (20

*La procédure de codécision est le procédé (parfois long) qui permet au Parlement et au Conseil européens de valider des actes législatifs de
Maastricht en 1993, elle a été simplifiée et son champ élargi par les traités d'Amsterdam (1999) et de Nice (2003).*

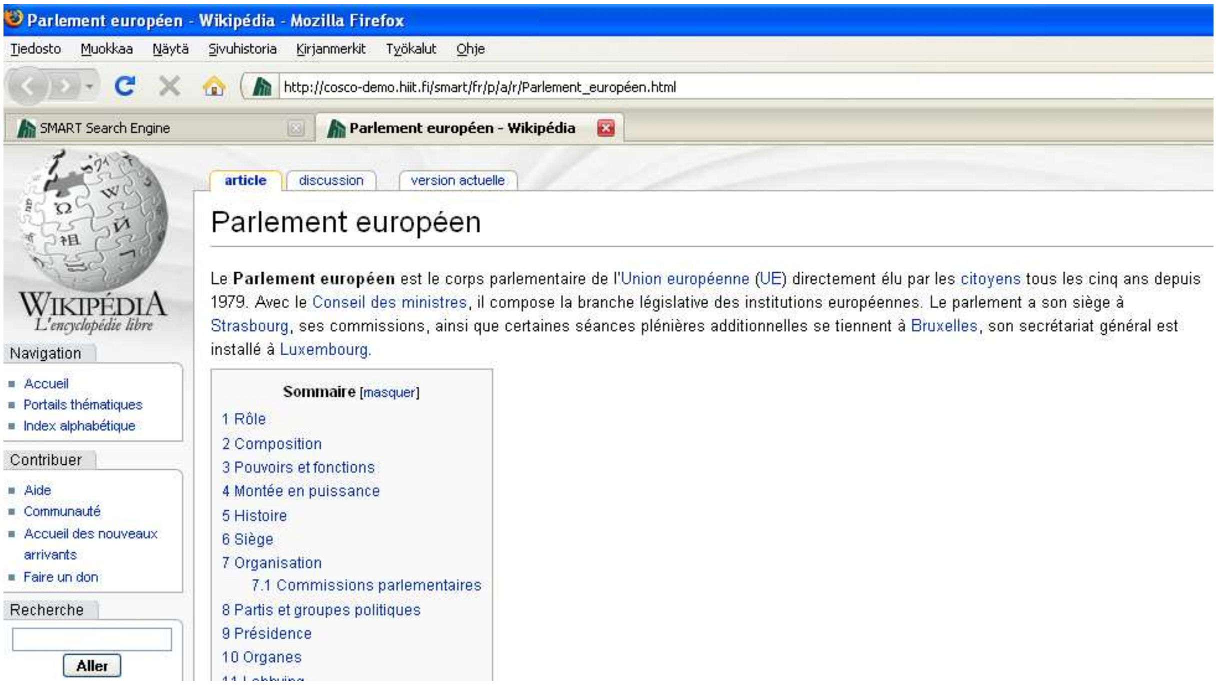
European Council

Conseil européen

The European Council is the Summit of the leaders of the European Commission, and Heads of State or government leaders of the 2
set up in December 1974 in Paris at the Summit of the European Economic Community (EEC), on the initiative of President Giscard d

*Le Conseil européen désigne le sommet des dirigeants de la commission européenne, et chefs d'État ou chefs de gouvernement des vingt
informel créé en décembre 1974 à Paris lors d'un sommet de la Communauté économique européenne (CEE), à l'initiative du Président Gis*

Figure 2.2: Results of the query in the French Wikipedia.



http://cosco-demo.hiit.fi/smart/fr/p/a/r/Parlement_européen.html

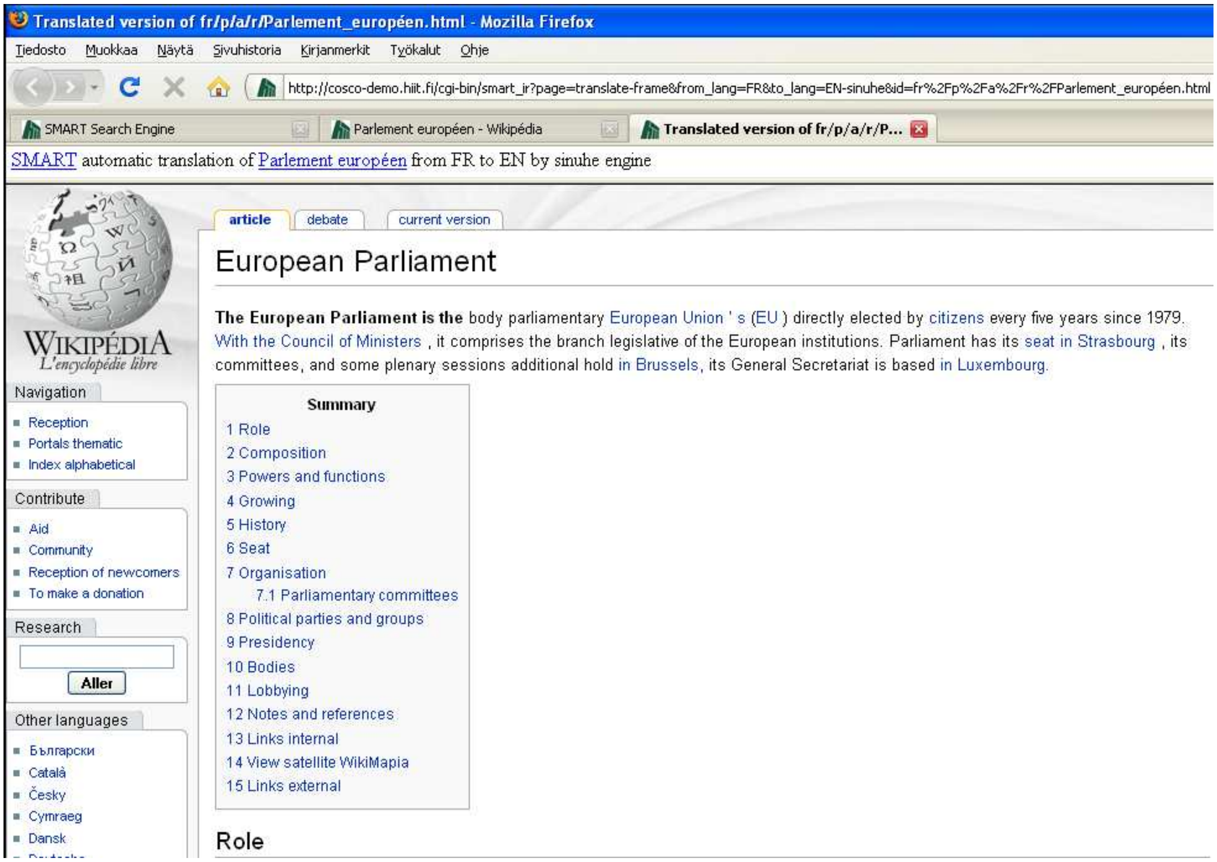
Parlement européen

Le **Parlement européen** est le corps parlementaire de l'Union européenne (UE) directement élu par les citoyens tous les cinq ans depuis 1979. Avec le Conseil des ministres, il compose la branche législative des institutions européennes. Le parlement a son siège à Strasbourg, ses commissions, ainsi que certaines séances plénières additionnelles se tiennent à Bruxelles, son secrétariat général est installé à Luxembourg.

Sommaire [masquer]

1 Rôle
2 Composition
3 Pouvoirs et fonctions
4 Montée en puissance
5 Histoire
6 Siège
7 Organisation
7.1 Commissions parlementaires
8 Partis et groupes politiques
9 Présidence
10 Organes

Figure 2.3: The original version.



SMART automatic translation of Parlement européen from FR to EN by sinuhe engine

European Parliament

The European Parliament is the body parliamentary European Union ' s (EU) directly elected by citizens every five years since 1979. With the Council of Ministers , it comprises the branch legislative of the European institutions. Parliament has its seat in Strasbourg , its committees, and some plenary sessions additional hold in Brussels, its General Secretariat is based in Luxembourg.

Summary

1 Role
2 Composition
3 Powers and functions
4 Growing
5 History
6 Seat
7 Organisation
    7.1 Parliamentary committees
8 Political parties and groups
9 Presidency
10 Bodies
11 Lobbying
12 Notes and references
13 Links internal
14 View satellite WikiMapia
15 Links external

Role

Figure 2.4: The article in French Wikipedia translated to English.

Chapter 3

Communication protocol

See Fig. 1.1 for an overview of the architecture. We will here describe the way the modules communicate.

3.1 CLIR module interface

An example request to the CLIR engine, using our `evropski parlament` example:

```
<?xml version="1.0" encoding="UTF-8"?>
<clirRequest>
 <queryId>some_id</queryId>
 <queryLanguage>SL</queryLanguage>
 <targetCollection>FR Wikipedia</targetCollection>
 <engine>xerox</engine>
 <startDocNum>1</startDocNum>
 <maxDocNum>10</maxDocNum>
 <query>evropski parlament</query>
</clirRequest>
```

Note that the engine can be specified using the `engine` element. The options are: `xerox`.

An example reply from the CLIR engine:

```
<?xml version="1.0" encoding="UTF-8"?>
<clirReply>
 <queryId>some_id</queryId>
 <startDocNum>1</startDocNum>
 <endDocNum>10</endDocNum>
 <relevantDocs n="10">
  <doc score="-2.95192">fr/p/a/r/Parlement.html</doc>
  <doc score="-3.42988">fr/p/a/r/Parlement_de_Nouvelle-
ZÃ©lande_589e.html</doc>
  <doc score="-3.54169">fr/p/a/r/Parlement_ottoman.html</doc>
  <doc score="-3.55593">fr/p/a/r/Parlement_europÃ©en.html</doc>
  <doc score="-3.62946">fr/p/a/r/Parlement_flamand.html</doc>
  <doc score="-
3.64205">fr/p/a/r/Parlement_Ã©tudiant_du_QuÃ©bec_3378.html</doc>
  <doc score="-3.6432">fr/p/a/r/Parlement_du_Canada_62d3.html</doc>
  <doc score="-3.67453">
fr/l/i/s/Liste_des_105_membres_de_la_Convention_sur_l’avenir_de_l’Europe_df3
6.html
  </doc>
  <doc score="-
3.73491">fr/a/s/s/AssemblÃ©e_nationale_du_QuÃ©bec_2778.html</doc>
  <doc score="-3.73721">fr/p/a/r/Parlement_de_Normandie_f53e.html</doc>
 </relevantDocs>
</clirReply>
```

Note that all that is required of the score is that it is numerical and enables ranking the hits.

3.2 SMT module interface

An example request to the SMT server:

```
<?xml version="1.0" encoding="UTF-8"?>
<translation_request>
  <source_language>FR</source_language>
  <target_language>EN</target_language>
  <speed>fast</speed>
  <translation_engine>Moses</translation_engine>
  <source>
    <s ID="1">Sentence in Fr 1</s>
    <s ID="2">Sentence in Fr 2</s>
    <title ID="3">Some title in Fr</title>
    <li ID="4">Some list item in Fr</li>
  </source>
</translation_request>
```

Language codes are:

- FR for French
- ES for Spanish
- SL for Slovenian
- EN for English

IDs are necessary to align source and target sentences in the reply. Titles are marked apart as they might require a slightly different translation model.

An example reply from the SMT server: (<trs> contains a recased detokenized translated sentence)

```
<?xml version="1.0" encoding="UTF-8"?>
<translation_reply>
  <source>
    <s ID="1">
      <w ID="1.1">il</w>
      <w ID="1.2">s\’</w>
      <w ID="1.3">agit</w>
      <w ID="1.4">d\’</w>
      <w ID="1.5">une</w>
      <w ID="1.6">phrase</w>
      <w ID="1.7">.</w>
    </s>
  </source>
  <translation>
    <s REF="1">
      <w REF="1.1 1.7">It is a</w>
      <w REF="1.8 1.8">sentence</w>
      <w REF="1.9 1.9">.</w>
    </s>
    <trs REF="1">It is a sentence.</trs>
  </translation>
</translation_reply>
```

Note how the REF element marks the alignment.

Chapter 4

Module implementation details

4.1 Graphical Use Interface

4.1.1 Univ. of Helsinki implementation

4.2 CLIR module

4.2.1 Xerox implementation

4.2.2 Univ. of Helsinki implementation

4.3 HTML translation module

4.3.1 Amebis implementation

4.3.2 Univ. of Helsinki implementation

4.4 SMT module

4.4.1 Moses/Sinuhe Univ. of Helsinki implementation

4.4.2 SOTON implementation

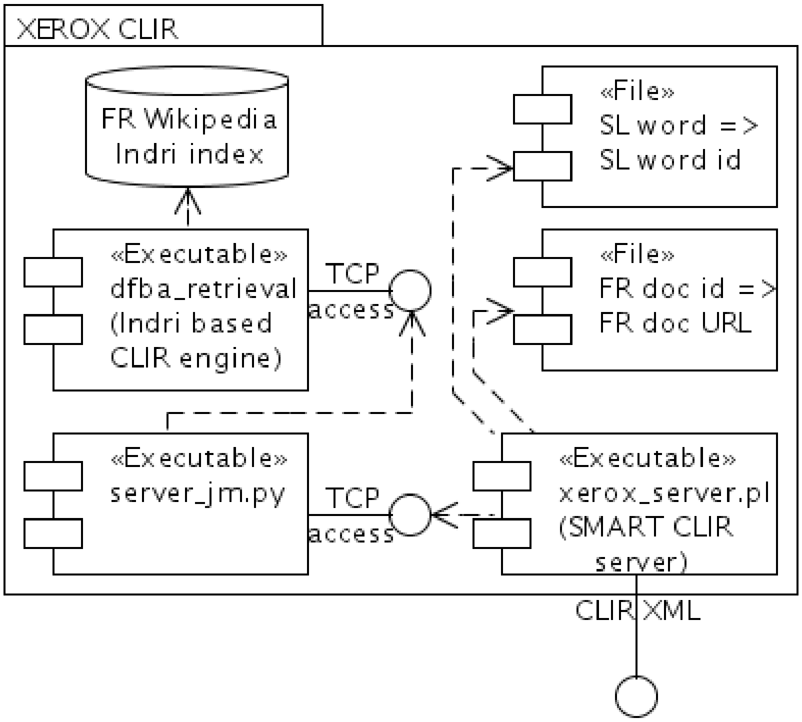

XEROX CLIR
FR Wikipedia Indri index
«File» SL word => SL word id
«Executable» dfba_retrieval (Indri based CLIR engine)
TCP access
«File» FR doc id => FR doc URL
«Executable» server_jm.py
TCP access
«Executable» xerox_server.pl (SMART CLIR server)
CLIR XML

Figure 4.1: The architecture wrapped around the Xerox CLIR engine.

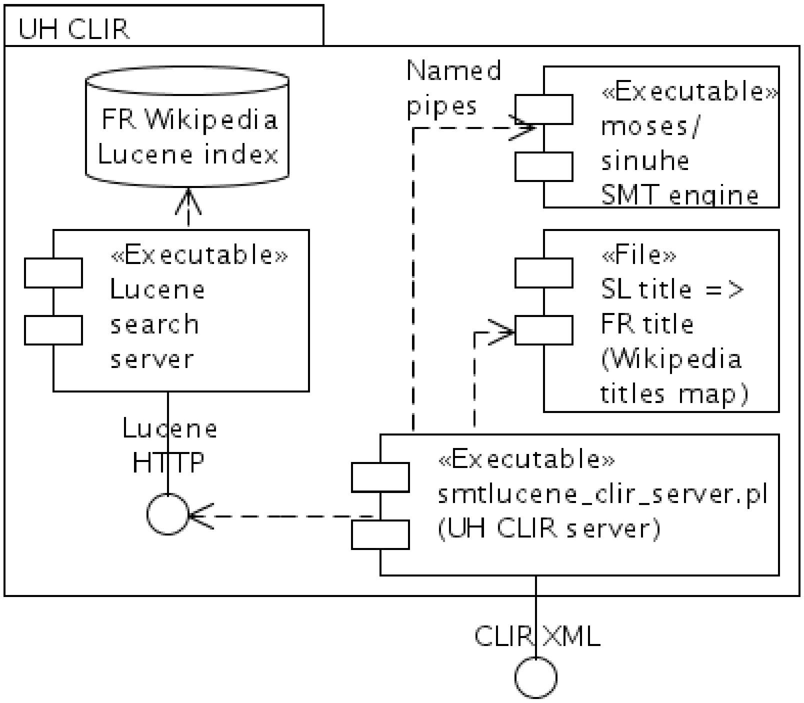

UH CLIR
FR Wikipedia Lucene index
Named pipes
«Executable» moses/ sinuhe SMT engine
«Executable» Lucene search server
«File» SL title => FR title (Wikipedia titles map)
Lucene HTTP
«Executable» smtlucene_clir_server.pl (UH CLIR server)
CLIR XML

Figure 4.2: The architecture of the Univ. of Helsinki CLIR engine.

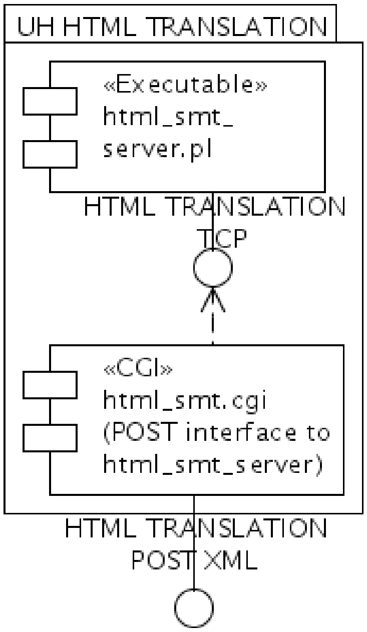

UH HTML TRANSLATION
«Executable» html_smt_ server.pl
HTML TRANSLATION TCP
«CGI» html_smt.cgi (POST interface to html_smt_server)
HTML TRANSLATION POST XML

Figure 4.3: The architecture of the Univ. of Helsinki HTML translation engine.

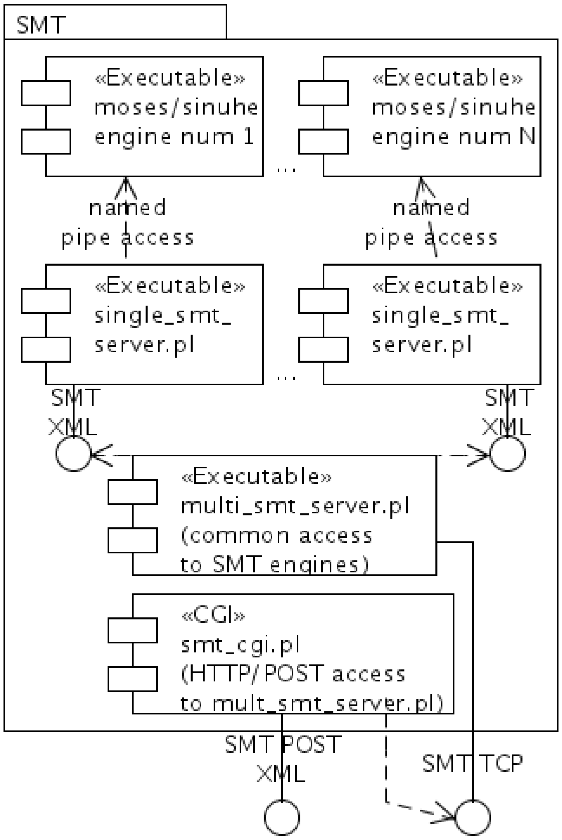


Figure 4.4: The architecture of the Univ. of Helsinki machine translation engine.

Annex