L&H Approach
Towards New Languages

Gr. Thurmair, J. Ritzke
EAMT
Prague 4/1999
Two Level Approach

- Integration of existing systems
  - the iTranslator platform
  -> to be able to offer these languages

- Development of new components
  - components and applications
  -> to have a solid and modular integration
1 System Integration with iTranslator

- **Purpose of iTranslator**
  - Support Translation Requests via Internet

- **Method**
  - Platform for integrating multivendor systems
  - API’s for clients and for engines

- **Transparent Access**
  - uniform external access for users

- **non-ascii & multibyte charset support**
iTranslator: multi-vendor platform

- **integrated systems:**
  - T1 (En, De, Fr, Es) (on engine level)
  - Globalink (En, De, Fr, Es, Po, It)
  - AppTek (Ar, Ko)
  - AILogic (Ja)

- **link to other systems tested**
  - Logos (on client level)
  - PaTrans (LE-OTELLO project)

- **additional systems in progress**
iTranslator Architecture

Client: launch jobs
check status

iT Master: schedule jobs
administer user DB accounting

Client API

(scheduler) Logos

MT-Engine (Logos)

Engine API

MT-Engine execute jobs (Globalink)

MT-Engine execute jobs (T1)
iTranslator User Interface

The iTranslator software interface is shown on the screen. It includes options for translating documents, selecting translation services, and specifying the direction and type of translation. The interface also allows for the selection of general vocabulary and subject areas, and provides options for result retrieval via email or download.
available

- English <> German
- English <> Spanish
- English <> French
- English <> Italian
- English <> Portuguese
- English <> Japanese
- English <> Arabic
- English <> Korean
- English > Chinese

in progress

- German <> French
- German <> Italian
- Italian <> French
- Italian <> Spanish
- Japanese <> Chinese
- Russian > German
- English <> Dutch
- English <> Russian
- German > Russian
Integration of new language pairs

- **Support Engine API for integration**
  - language pair
  - file to be translated
  - subject area
  - (memory modules)

- **Administration Logistics**
  - lexicon administration on server side
  - OLIF lexicon exchange format (LE-OTELO)
2 Development of new components

Development steps
- Linguistic Definitions
- Lexicon Development
  - monolingual, bilingual resources
- Morphological Components
  - analysers (lemmatisers)
  - generators (flexers)
- Partial syntactic analysis
  - special purpose recognisers, taggers
- Syntactic analysis
- Transfer components
Development of New Components (2)

- **Purpose**
  - support of L&H business lines
    - Translation (from term level to full MT)
    - information Retrieval & Extraction
    - (Speech)

- **Approach**
  - step - by - step development
  - integration of partial results into applications

- **Languages**
  - Russian, Serbian, Croatian, Albanian, (others)
Linguistic Definitions

- Tagset
  - standard tagset
  - extended tagset (incl. morphology codes)
- Lexical and grammatical features

Standardisation to other language pairs
- (subject area hierarchy etc.)

Software-Integration issues
- character code, converters, ...
Lexicon Development

- **Function Word Dictionary**
  - collects closed word classes

- **Open Word Classes**
  - general vocabulary
  - special subject areas for vertical domains

- **Special Word Classes**
  - proper name analysis
  - gazetteers for Information Extraction
Lexicon Development (2)

- Administration: Lexicon Database
  - Unicode back-end
  - project specific entry structures
  - common import/export format (OLIF)

- Application: Lexicon Lookup
  - internet lexicon lookup for end users
Lexicon Lookup: Example Interface
Lexicon Development Strategies

- license material
  - ELRA
  - other sources
- subcontract lexicon production
- corpus material to create input word lists
- existing material from the lexicon DB
- (mixture of all)

- (correction and quality control)
Morphological Components

- Single Word analysis and generation
  - learning components based on lexicons
    - Lemmatisers (input: text form)
      - standard tagset: base form, tag
      - extended tagset: base form, tag, morphology
    - Flexers (input: base form)
      - general: generate all possible forms (e.g. for IR)
      - dedicated: generate a specific form (e.g. in MT)

- Compilers to build specific data structures
Morphological Components (2)

- **Target Quality**
  - error rates < 4%

- **Possible Applications**
  - term extraction
  - information retrieval
    - query expansion
  - term translation and lexicon lookup

- **Basis for multiword and phrase analysis**
Special Pattern Recognition

- (multiword) terminology
  - purposes
    - lexicon lookup (known and unknown terms)
    - term extraction, topic identification
  - frequency and linguistic analysis

- information extraction
  - recognition of proper names
    - persons, places, ...

- finite state partial grammars / taggers
Example: Topic Identification

информация: Благодаря розыскам и informer установилось, что главный обвиняемый по названию Igor Volkov изготовляет amphetamine и после этого, с помощью carrier, sell этт drug разным customers. Сегодня приняли меры против лиц продавающих в промышленном предприятием бидоны с alcoholic beverages container. В течение house search обыска нашлись в t. ч. chemicals, которые можно использовать для production amphetamine. В закрытом помещении находились посудина с ручным электромиксером и с погруженным электрокипательником, которые были использованы для drug production. Кроме заказчика Igor Volkov действовали два мужчина в качестве изготовителя. Дальнейшие домашние обыски проводили до сохранения 10 kilogram amphetamine. Заказчик Игорь Волков вероятно убежал из Russia - поэтому мы просим вас начать розыск за него.
Possible Applications

• Named Entity Analysis
• multilingual Information Retrieval
  • multiword term translation
• topic identification and clustering
• term translation
• terminology tools
  • e.g. proof-reading
Example: Information Extraction

Corpus of C:\Program Files\LHS\Aventinus\texts\namertexts\en-101.txt

(1) Colombia far from ending drug traffic, official says RTw 6/23/95 6:23 PM
Copyright 1995 Reuters Ltd. All rights reserved. (2) The following news
report may not be republished or redistributed, in whole or in part, without the
prior written consent of Reuters Ltd.

(3) BOGOTA , June 25 ( Reuter ) - Colombia 's government might be close to
defeating the powerful Cali cartel with the capture and surrender of three of its
leaders, but it still has a long way to go before eliminating drug trafficking,
Defence Minister Fernando Botero said.

(4) Botero told the daily El Espectador that smaller drug cartels based in other
parts of the country were consolidating their positions in the trade while the
government was preoccupied with traffickers in the southwestern city of Cali.

(5) "There are organisations that have undoubtedly benefited a lot from the
concentration of efforts in Cali," he was quoted as saying in the Bogota
newspaper 's Sunday editions. (6) Drug trafficking is much more than the Cali
cartel .

(7) Even former members of the Medellin cartel, led by Pablo Escobar until his
death in 1993, were regrouping, he said.

(8) Colombia is the world 's biggest cocaine exporter and the third largest
heroin exporter, according to U.S. officials.

(9) Botero said the government was studying ways to fight the smaller cartels
by creating special army and police teams similar to the one pursuing the Cali
cartel, which controls 80 percent of the cocaine smuggled into the United
Syntactic Analysis

- Different levels of complexity
  - query analysis in Information Retrieval
  - Dependency structures for IR
  - Phrase structures for MT

- Grammar formalism of METAL/T1
  - augmented transformational approach
  - special development environment

- Applications
  - Information Extraction / Scenario Analysis
  - Machine Translation
Result: Development Strategy

- application based development strategy
  - verify every step in direct applications
    - lexicon in lexicon lookup components
    - morphology in term and retrieval components
    - etc.
  - frequency and linguistic analysis

- reduce “time to market”
  - do not wait 2 years until the miracle is born
  - produce tangible results asap

- re-use software components & platforms
Status

- Lexicon work in progress
  - for several Middle and East European languages
  - between 5000 and 60000 entries
    - monolingual, bilingual
    - some special entries, special domains

- Morphology components in progress

- Partial grammars in progress

- Underlying software in place