

L&H Approach Towards New Languages

Gr. Thurmair, J. Ritzke

EAMT

Prague 4/1999



Gesellschaft für
Multilinguale Systeme

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

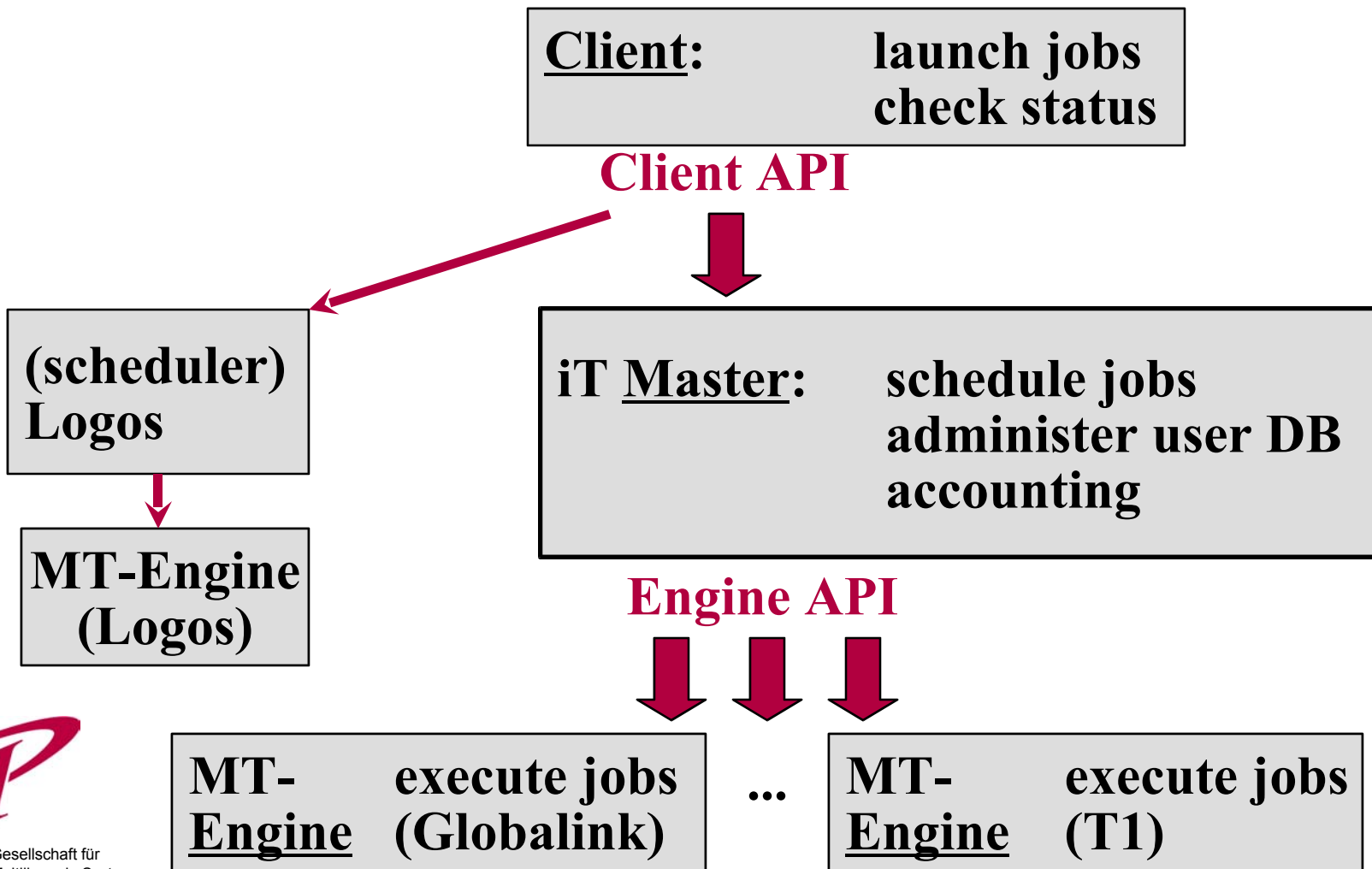
iTranslator: multi-vendor platform

- **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



iTranslator User Interface

The screenshot displays the iTranslator Net application window. The main window shows a list of files with columns for 'Service' and 'Document name'. A 'Translate Document' dialog box is open in the foreground, allowing the user to select a document, specify the translation direction, choose a job type, and select subject areas.

Main Window:

Service	Document name
Standard translation	Active Setup Log.txt
Standard translation	Markt.rtf
Cost estimation	Wald.rtf
Standard translation	test1.rtf
Standard translation	Leather.rtf
Gold translation	Umweltverschmutzung.RTF
Standard translation	about.html
Standard translation	related.html
Standard translation	related.html
Standard translation	about.html
Standard translation	about.html

Translate Document Dialog:

- Document: C:\Demo\Barcelona Tech\iTranslator Net\Germ_Fren.r
- Translation direction: English-German
- Job type: Machine Translation (MT) Post-Edited MT Human Translation
- Subject areas: General Vocabulary (selected), Common Social Vocabulary, Art & Literature, Ecology, Economy & Trade, Government, International Relations, Law & Legal Science, Recreation, Sports, Games & Hobbies, Social Science, Common Technical Vocabulary, Agriculture & Fishing, Civil Engineering
- Result retrieval: Email Download
- Email address: zeljko.angjelkoski@lhs-It.de

Buttons: Cost, Translate, Cancel

iTranslator Languages

available

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

- German <> French
- German <> Italian
- Italian <> French
- Italian <> Spanish
- Japanese <> Chinese
- Russian > German

in progress

- 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-OTELLO)

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

■ 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

Netscape
File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Stop

Location: [http://muc112-pc.gsmuc.de:80/db2muc_weblexdb/plsql/wgms\\$login.errdblogin](http://muc112-pc.gsmuc.de:80/db2muc_weblexdb/plsql/wgms$login.errdblogin) What's Related

Instant Message WebMail Contact People Yellow Pages Download Channels

English Term

ID **15** Canonical form:

Parts of speech(POS): Subject area:

Grammar: Usage:

Entry type: Source:

Definition: Comment:

French/German/Spanish Term

French equivalent: French POS:

German equivalent: German POS:

Spanish equivalent: Spanish POS:

Built today: 26.11.1998 08:05
 1998 ©L&H Language Technology Division München

Central Entry - Query Result

ID	Canonical form	Pos	Subj. area
4	USA	noun	DP-SW
5	heroin	noun	GV
6	cocaine	noun	GV
8	Heroin	noun	GV
9	LSD	noun	DP
10	Crack	noun	DP
11	Hashish	noun	DP
12	Pot	noun	DP
13	cannibis	noun	DP
14	Phrenology	noun	DP
15	Shirlock Holmes	noun	DP
16	Felon	noun	DP

12 of 32 from Total: (34)

Built today: 26.11.1998 08:05
 1998 ©L&H Language Technology Division München

Create by: Alric Jackson, Date created= 23-NOV-98, Modified by: Alric Jackson, Date modified= 25-NOV-98

Lexicon Development (3)

■ 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

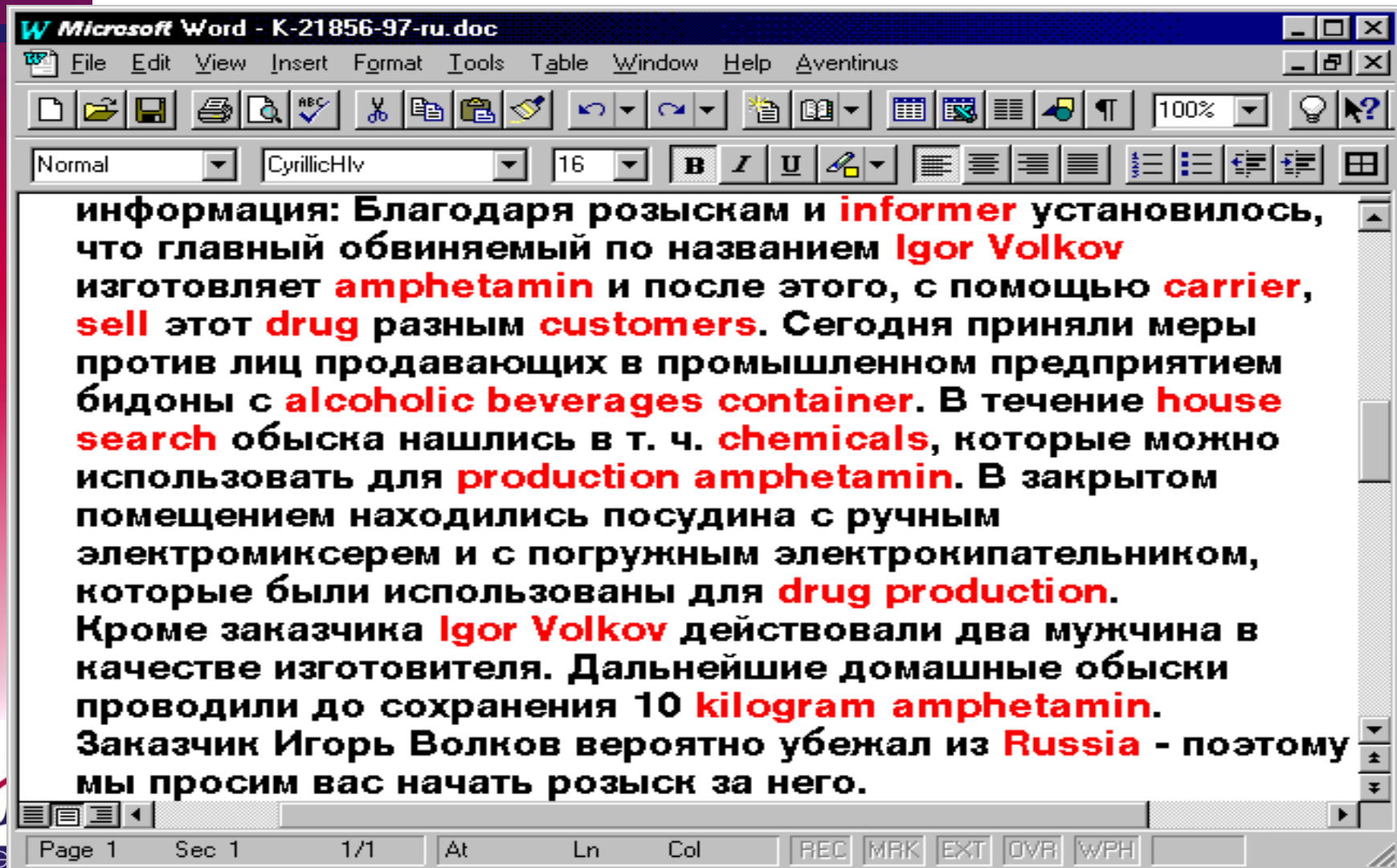
■ information extraction

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

■ finite state partial grammars / taggers



Example: Topic Identification



Microsoft Word - K-21856-97-ru.doc

File Edit View Insert Format Tools Table Window Help Aventinus

Normal CyrillicHlv 16 B I U

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

Page 1 Sec 1 1/1 At Ln Col REC MRK EXT OVR WPH

Special Pattern Recognition (2)

■ 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

Report Report0005_EN - Microsoft Internet Explorer

File Edit View Go Favorites Help

Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Print Edit

Address C:\gregor\NE\Reports\EN\Report0005_EN_frm.html Links

Type	Term	freq.	sentence(s)
Company	Reuters Ltd.	2	1 , 2
Drug	cocaine	2	8 , 9
Drug	drug	4	1 , 3 , 4 , 11
Drug	heroin	1	8
Organisation	Cali cartel	5	3 , 6 , 9 , 10 , 12
Organisation	Colombian government	1	10
Organisation	Medellin cartel	1	7
Organisation	REUTER	1	15
Organisation	Reuter	1	3
Person	Botero	2	4 , 9
Person	Pablo Escobar	1	7
Person	Fernando Botero	1	3
Person	Myles Frechette	1	10
Person	Victor Julio Patino Fomeque	1	12
Person	Gilberto Rodriguez Orejuela	1	14
Place	BOGOTA	1	3
Place	Bogota	1	5
Place	Cali	3	4 , 5 , 13

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

(1) [Colombia](#) far from ending [drug](#) traffic , official says RTw 6/25/95 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](#)

My Computer

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**