linguatec Sprachtechnologien

Crosslingual Information Retrieval (CLIR)

Gr. Thurmair
Pisa 2004

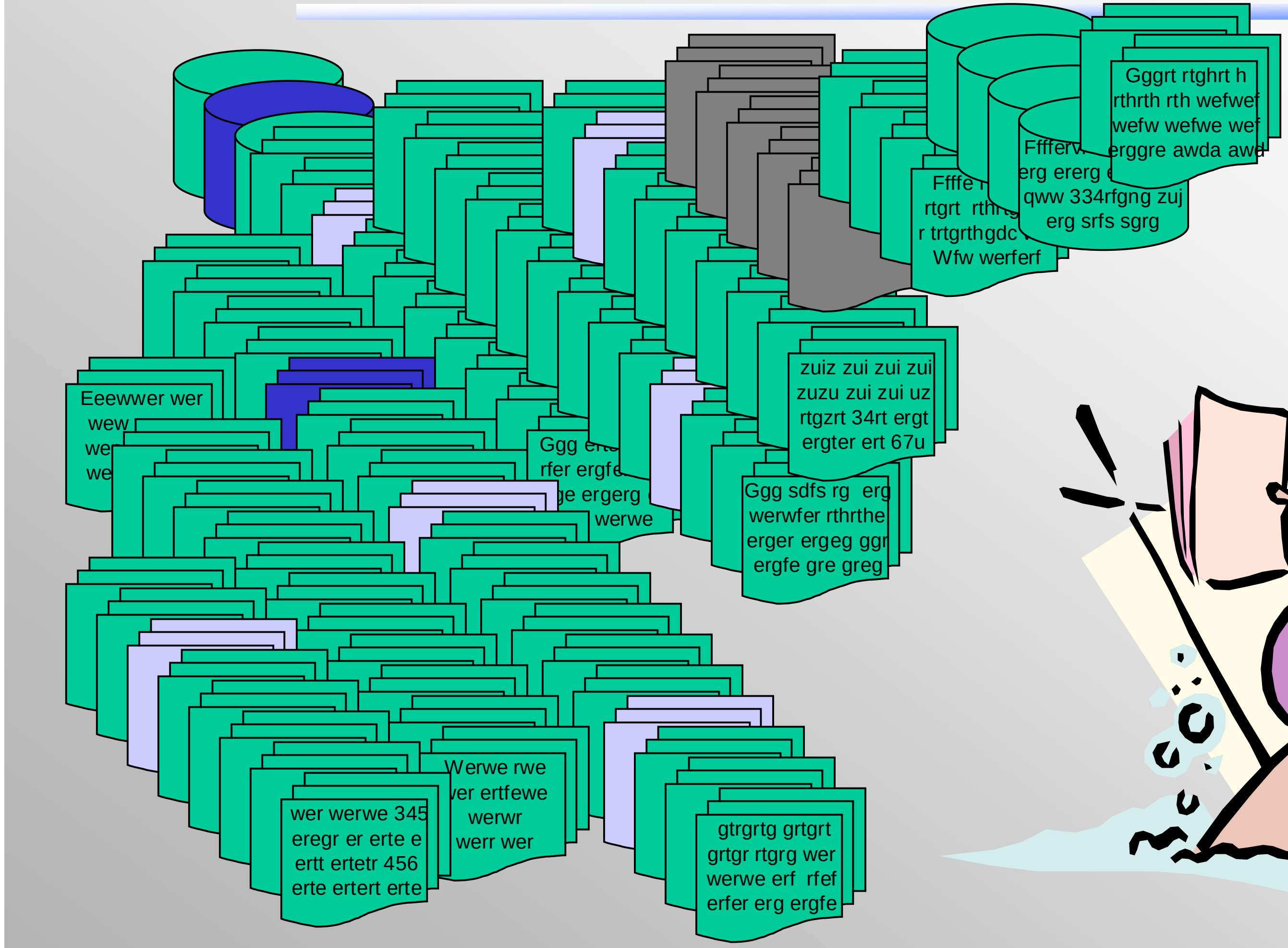
Outline

- Current retrieval
  - Likely improvements
- Why this does not easily work
- Design of a CLIR
  - Text translation, index translation, query translation
- Architecture of QTE based system
  - Analysis
  - Index
  - Query translation
  - Search & ranking
  - Document retranslation

Outline

I From IR to CLIR

Looking for trouble?



previous ◀ ▶ next

Architecture of an Indexing System

- Example: Lucene Open Source IR System
  - www.apache.org/lucene
  - Current Version = V1.4

- Software Modules:
  - Document
    - Document maintenance (set of fields)
  - Analysis
    - Tokeniser, Stop list, Index term Generator
  - Index
    - Writer: create / add documents
    - Reader: access index data
  - Search
    - Query Parser
    - Queries (Phrases / Boolean / …
    - Searcher (Hits, Highlighting …)

What is the problem in search?

- Automatic Indexing: **From strings to concepts!**
  - Machines process *strings* – humans look for concepts
- Take into account variablitiy and flexibility of human language
  - Reduce variance / normalise to the relevant concepts of a text
- Finding relevant terms is a linguistic job
  - Linguistics = science of language
- Indexing is a linguistic task, among others with the support of statistic components

What is an index term?

XML Topic Maps: XTL V1.0 (2001)

> This specification provides a model and grammar for representing the structure of information resources used to define topics, and the associations (relationships) between topics. Names, resources, and relationships are said to be *characteristics* of abstract subjects, which are called *topics*. Topics have their characteristics within *scopes*: i.e. the limited contexts ... One or more interrelated documents employing this grammar is called a ”topic map.”
>
> ...
>
> A **<u>topic</u>** is a resource within the computer that stands in for (or ”reifies”) some real-world *subject*. Examples of such subjects might be the play *Hamlet*, the playwrighter ***William Shakespeare***, or the ”authorship” relationship.

=> is *<u>Shakespeare</u>* a relevant term?

(38 occurrences, one of top 50 Terms in the corpus)

Current Index Term creation techniques

- Tokenisation
  - All index terms are single words!
- Stemming
  - Intends to neutralise morphology variants
- Context
  - Intends to capture syntactic structure
- Clustering
  - Intends to neutralise semantic variants

author	
Author	*ing*
author	*ities*
author	*s*
Author	*s*
author	*ship*

a topic maps
several elements
into a commmon
cluster ...

topic element subject map XTM name association scope
ID document resource occurrence class role indicator reference
child variant relationship specification ... Shakespeare ...

Morphology: Stemmer

- **Stemming**: Reduction of Base forms to stems

  (often: Reduction of Textforms to stems)

- Technique: Cutting the **Affixes**
  - Problem: Cut prefixes as well?
  - Problem: Allomorphs
- Types of mistakes
  - Improper Linking of terms (Precision ↓)
  - Improper splitting of terms (Recall ↓)
- Effect
  - increases Recall
  - Increases noise if unprecise
  - Sometims necessary in CLIR (en verb - fr noun)

Verarbeitung	-> arbeit
Funktionalität	-> funkt
Integrierbarkeit	-> integr
überreagieren	-> reag
Überreaktion	-> reak?
Verdichtung	-> dicht
Versorgung	-> sorg
Regionalisierung	-> reg
Regierung	-> reg
Regen, regnerisch	-> reg

Search for ‚Topic Maps‘

- Indexed ‚terms‘ in the corpus are:

map	topic
Map	Topic
mapping	topic(s
Mapping	topicMap
maps	topicMaps.Org
map‘s	topicref
Maps	topics
	topic‘s
	Topics

- How must the query look like?

Result: TREC-8

Voorhees / Harmann 2000 (TREC8-Overview):

Standard Retrieval (ad-hoc task):

- No improvement in recent years
- Example: Precision SMART-System: between 0.15 and 0.4

„the results of the ad-hoc task have plateaued in recent years“

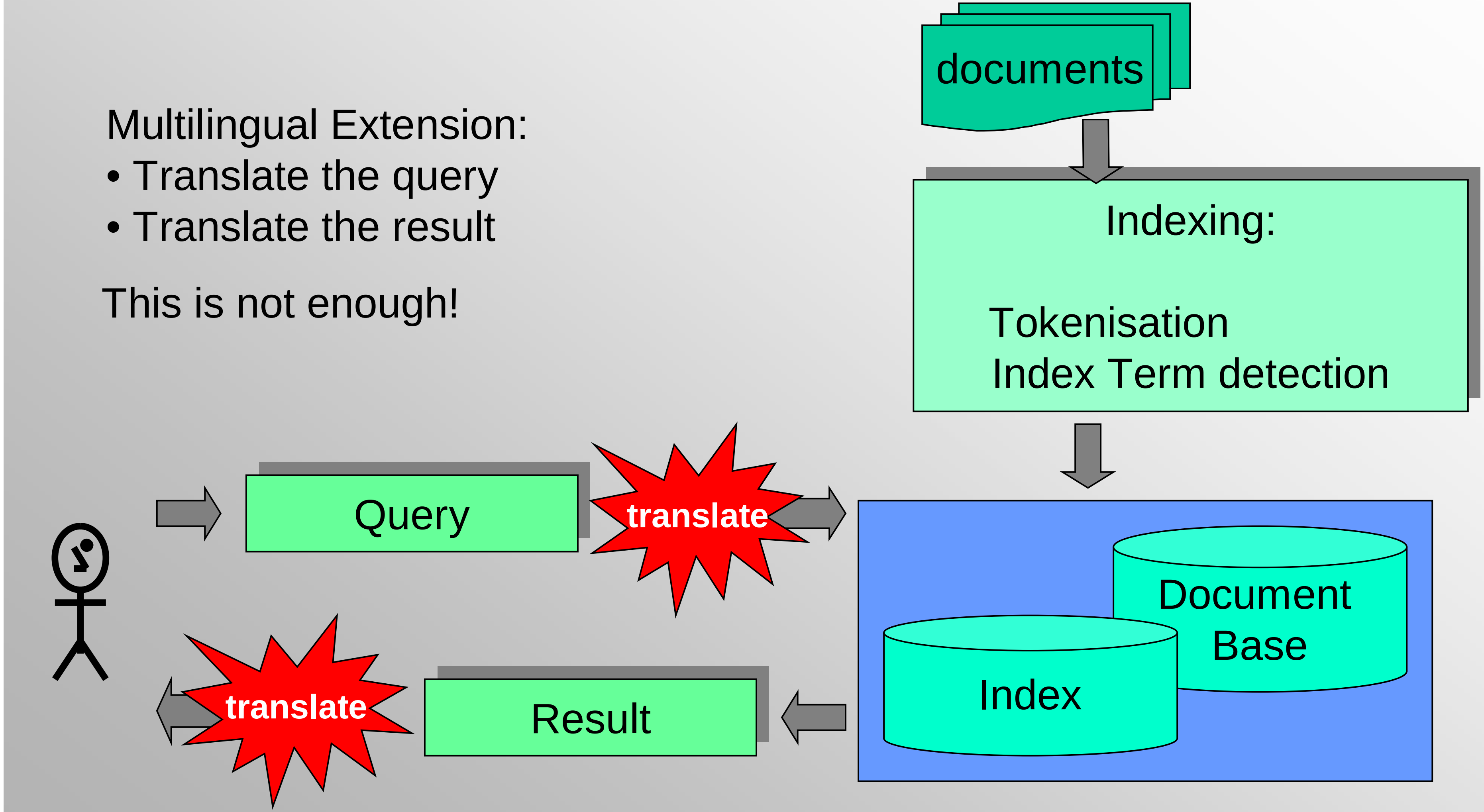
„the absolute performance of the task is less than ideal“

From IR to CLIR

Multilingual Extension:

- Translate the query
- Translate the result

This is not enough!



From IR to CLIR

- Current system structure is insufficient for good IR
  - Add more intelligence to the analysis
- Current structure also insufficient for CLIR
  - On index term creation
    - Index terms must be translated!
      - Single words, stemmed words … don‘t work
  - On index maintenance
    - Index must be language-sensitive …
  - On search
    - Search must be translated
    - Documents must be re-translated

From IR to CLIR: more Analysis Intelligence!

- Failure of standard procedures
  - **Truncation / Stemming**
    - Dictionaries do not contain translations for stems or text forms
    - *deal$* or *author$* are not translatable

    => Lemmatiser, Decomposer required
  - **Context** and context operators
    - Most terms to be translated are **Multiword terms**.
      Indizes contain only single words.
    - Translation of *money* NEAR *laundering*?

      De: no context operation as just 1 word (*Geldwäsche*)

      Es: wrong sequence (*lavado de dinero*)

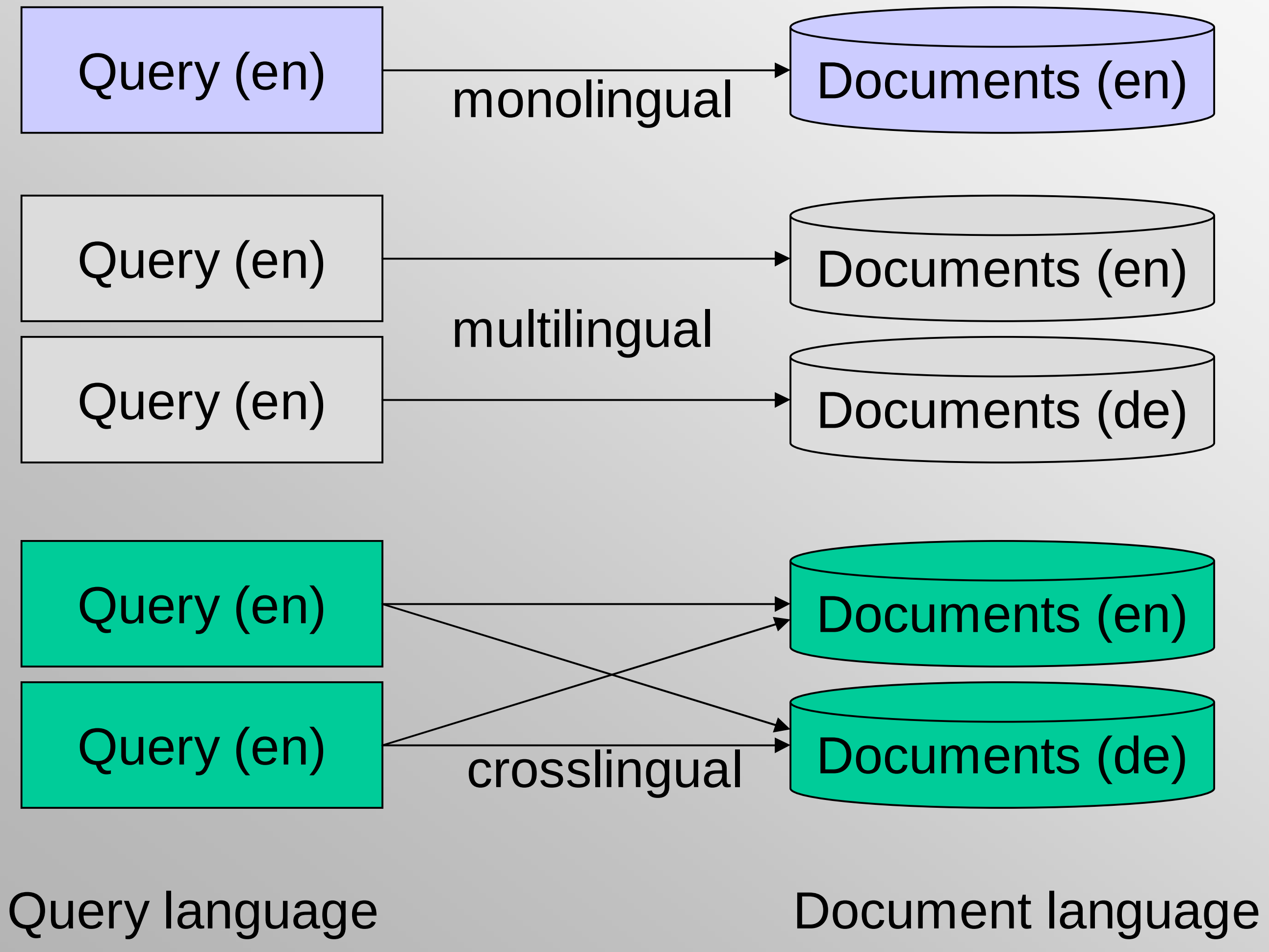
    => Dependency analysis requireed
  - Ranking / **Clustering**
    - Always relates to monolingual term universe
    - parallel Korpora if available do not really require CLIR

Outline

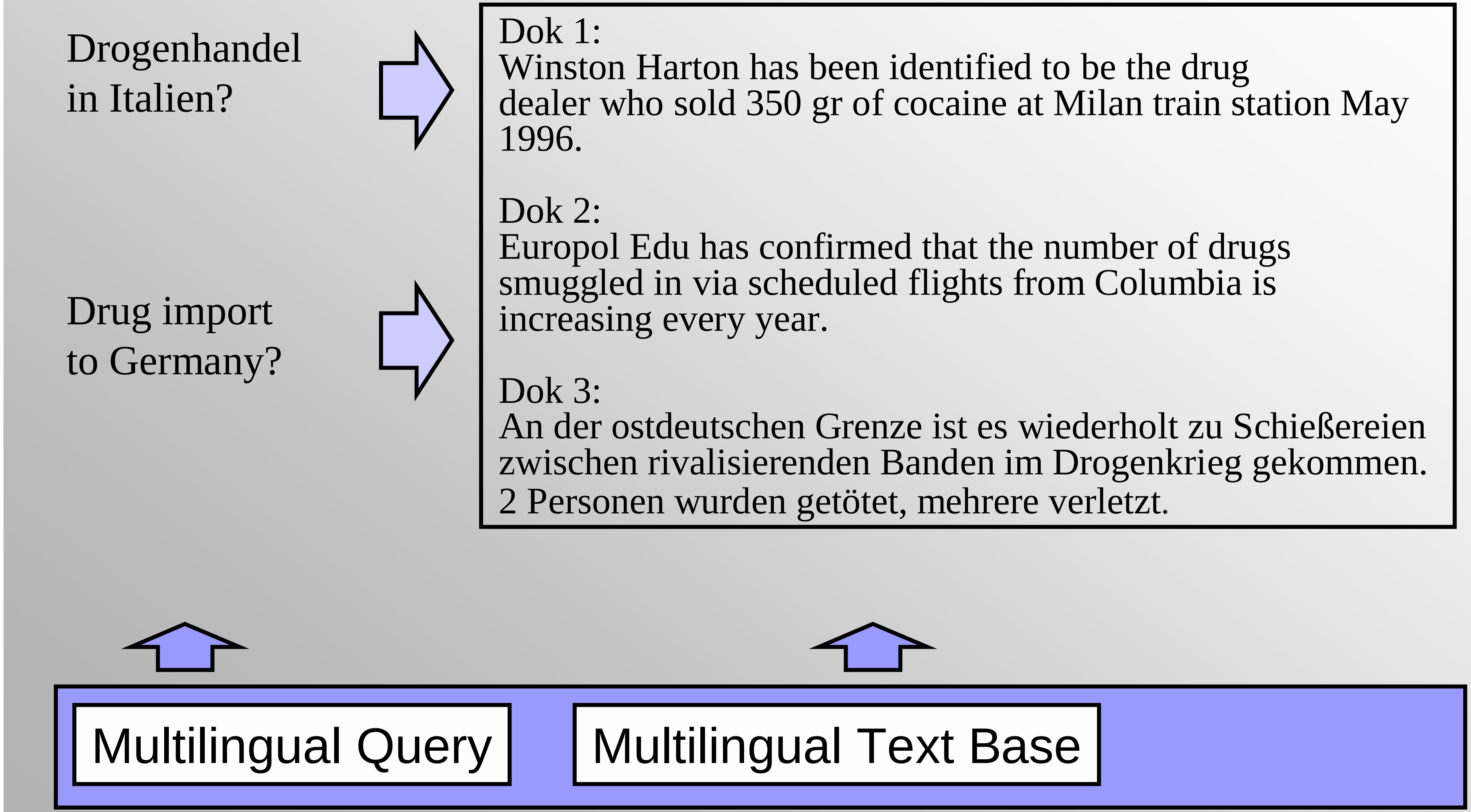
II CLIR Design

(terminology)

monolingual, multilingual, cross-lingual



Crosslingual Information Retrieval



Drogenhandel in Italien?

Drug import to Germany?

Dok 1:
Winston Harton has been identified to be the drug dealer who sold 350 gr of cocaine at Milan train station May 1996.

Dok 2:
Europol Edu has confirmed that the number of drugs smuggled in via scheduled flights from Columbia is increasing every year.

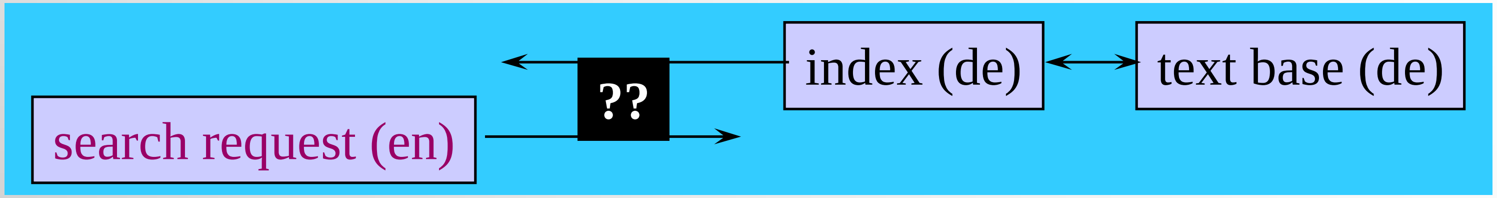
Dok 3:
An der ostdeutschen Grenze ist es wiederholt zu Schießereien zwischen rivalisierenden Banden im Drogenkrieg gekommen. 2 Personen wurden getötet, mehrere verletzt.

Multilingual Query

Multilingual Text Base

Architecture

Problem:



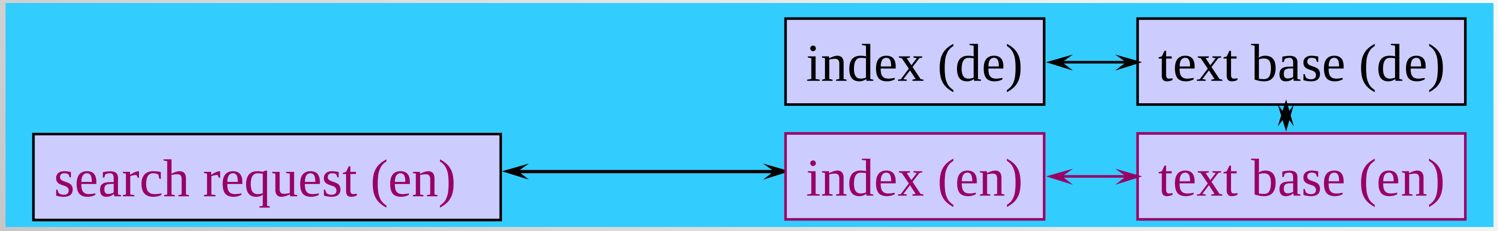
index (de)

text base (de)

??

search request (en)

1. **Translate Text:**



index (de)

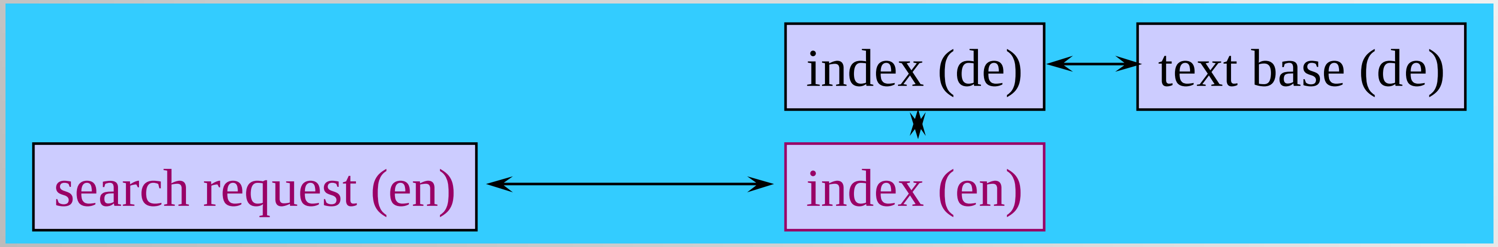
text base (de)

search request (en)

index (en)

text base (en)

2. Translate Index:



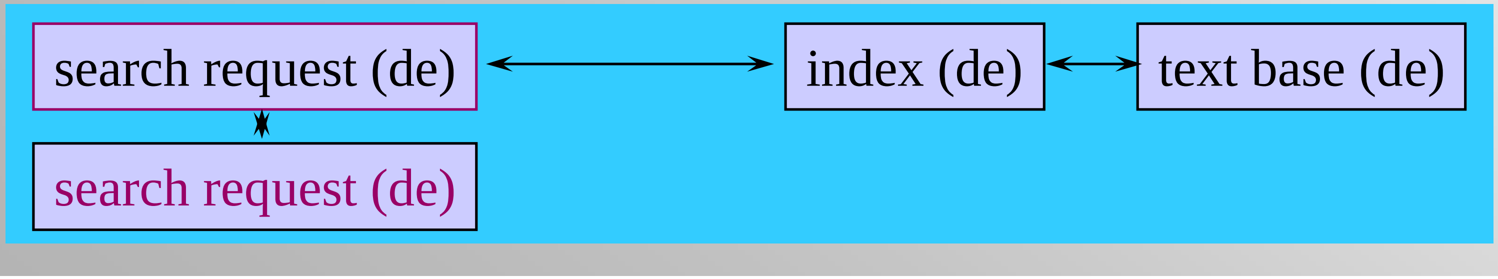
index (de)

text base (de)

search request (en)

index (en)

3. Translate Search:



search request (de)

index (de)

text base (de)

search request (de)

Multilingual Indexing

Options:

- Translate **Texts**
  - Case 1: into all relevant languages
    - Transforms crosslingual problem into monolingual one
  - Case 2: into Pivot language
    - Sub-case of above

- Translate **index terms**

- Translate **query**
  - Case 1: Query into n *mono*lingual document pools
  - Case 2: Query into a *multi*lingual document pool

Translation of texts 1

Drogenhandel
in Italien?

drug dealer
Drogenhändler
Drogenhandel
Mailand
Columbia
drug
drug gang
German

Drug import
to Germany?

Dok 1:
Winston Harton has been identified to be the drug dealer who sold 350 gr of cocaine at Milan train station May 1996.
Der **Drogenhändler**, der im Mai 1996 am **Mailänder** Bahnhof 350 Kokain verkauft hat, ist als Winston Harton identifiziert worden.

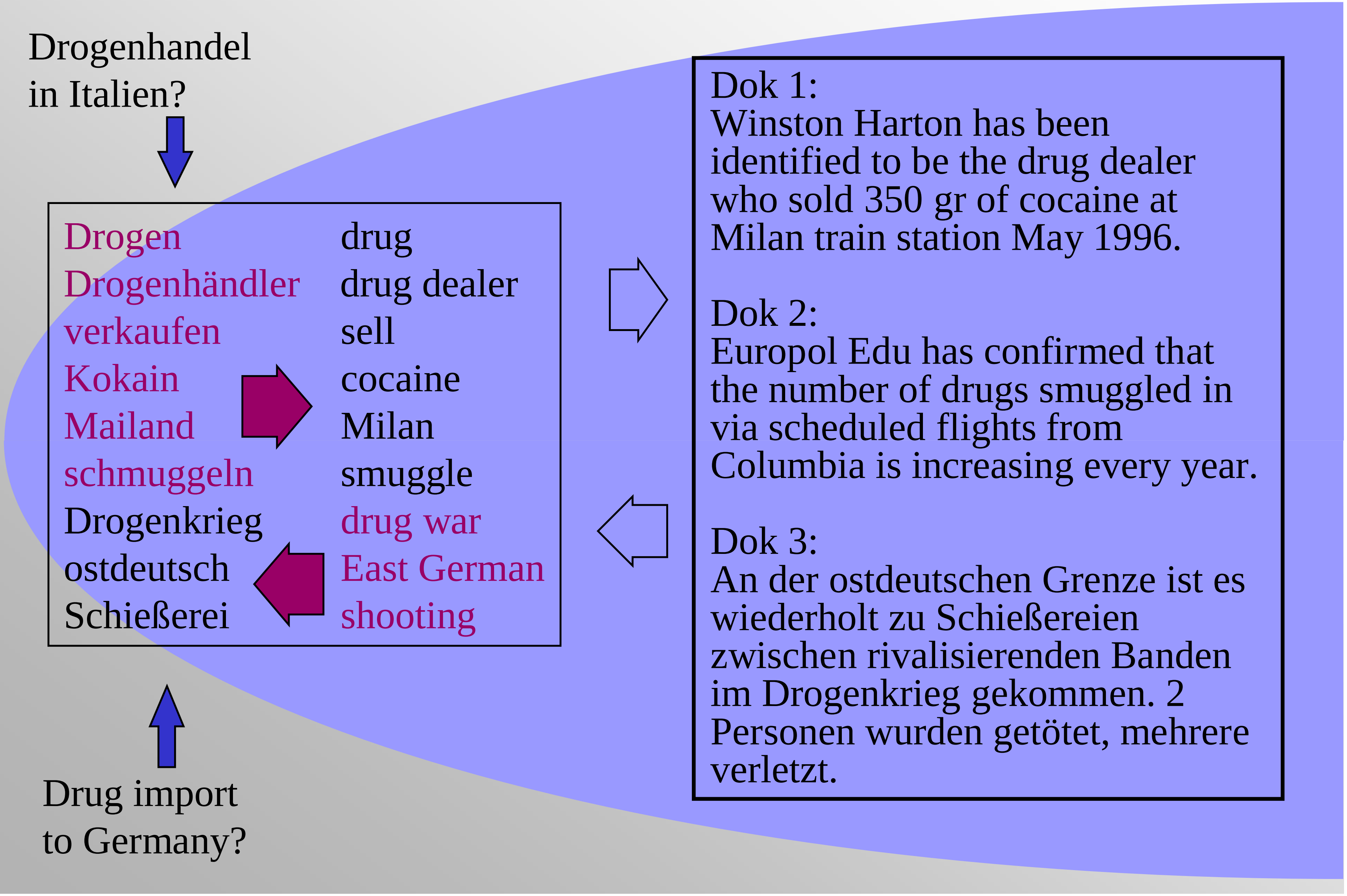
Dok 2:
Europol Edu has confirmed that the number of drugs smuggled in via scheduled flights from Columbia is increasing every year.
Europol EDU hat bestätigt, daß die Zahl der Drogen, die mit Linienflügen vom Kolumbien eingeschmuggelt werden, jährlich wächst.

Dok 3:
An der ostdeutschen Grenze ist es wiederholt zu Schießereien zwischen rivalisierenden Banden im Drogenkrieg gekommen. 2 Personen wurden getötet, mehrere verletzt.
There were repetitive shootings between rivalling **drug** gangs at the east **German** border. 2 people were killed, several others injured.

Translation of texts 2

- Problem solved by multiplying the texts
  - Make texts available in all languages
  - multilingual (= several monolingual) retrieval
- Feasibility:
  - Required in some applications
    - Patents, multilingual states (EG, Belgium, …)
  - Impossible in other areas (Internet)
- Evaluation:
  - From costly to impossible
  - Results depend on translation quality
    - translation dictionary updates invalidate search on existing document pool (->retranslate everything)

Translation of index 1



Drogenhandel
in Italien?

Drogen	drug
Drogenhändler	drug dealer
verkaufen	sell
Kokain	cocaine
Mailand	Milan
schmuggeln	smuggle
Drogenkrieg	drug war
ostdeutsch	East German
Schießerei	shooting

Drug import
to Germany?

Dok 1:
Winston Harton has been identified to be the drug dealer who sold 350 gr of cocaine at Milan train station May 1996.

Dok 2:
Europol Edu has confirmed that the number of drugs smuggled in via scheduled flights from Columbia is increasing every year.

Dok 3:
An der ostdeutschen Grenze ist es wiederholt zu Schießereien zwischen rivalisierenden Banden im Drogenkrieg gekommen. 2 Personen wurden getötet, mehrere verletzt.

Translation of index 2

- Idea:
  - – multilingual Index
    - Analyse query in query language, translate terms
    - Search with all document language index terms
  - – (Problem of retranslation of the hits)

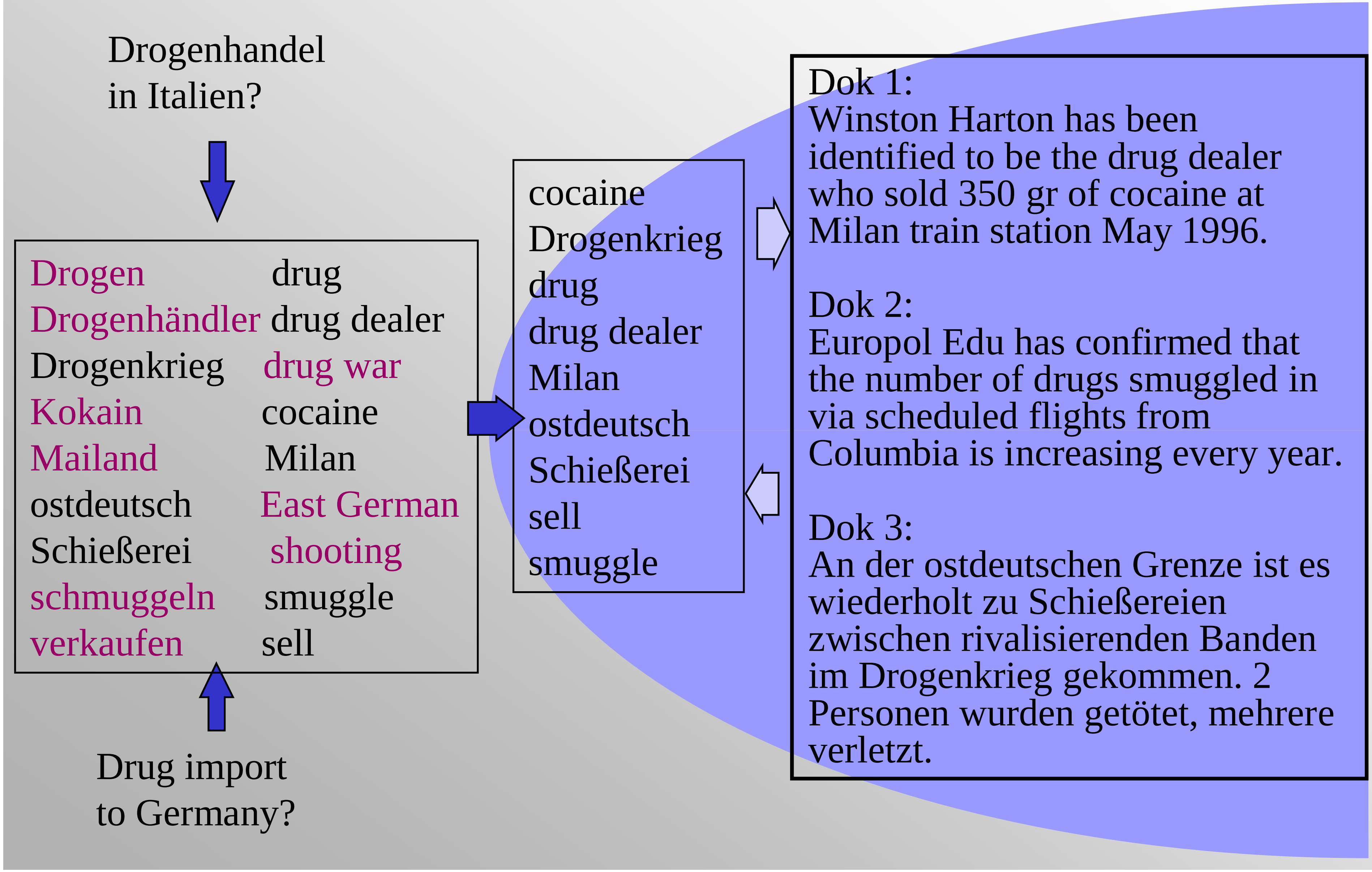
- Feasibility:
  - – Not feasible
    - Ambiguity of index terms
    - Multiword terms not in index
    - Context dependency of translations

=> Organise the index as a special resource!

Fehler:	*mistake, fault, error, bug*
nuclear:	*Kern~, zentral, nuklear*
power:	*Macht, Kraft, Strom*
plant:	*Pflanze, Unternehmen*

nuclear power plant:
zentrale Kraftpflanze?

Translation of query 1



Drogenhandel
in Italien?

Drogen	drug
Drogenhändler	drug dealer
Drogenkrieg	drug war
Kokain	cocaine
Mailand	Milan
ostdeutsch	East German
Schießerei	shooting
schmuggeln	smuggle
verkaufen	sell

Drug import
to Germany?

cocaine
Drogenkrieg
drug
drug dealer
Milan
ostdeutsch
Schießerei
sell
smuggle

Dok 1:
Winston Harton has been identified to be the drug dealer who sold 350 gr of cocaine at Milan train station May 1996.

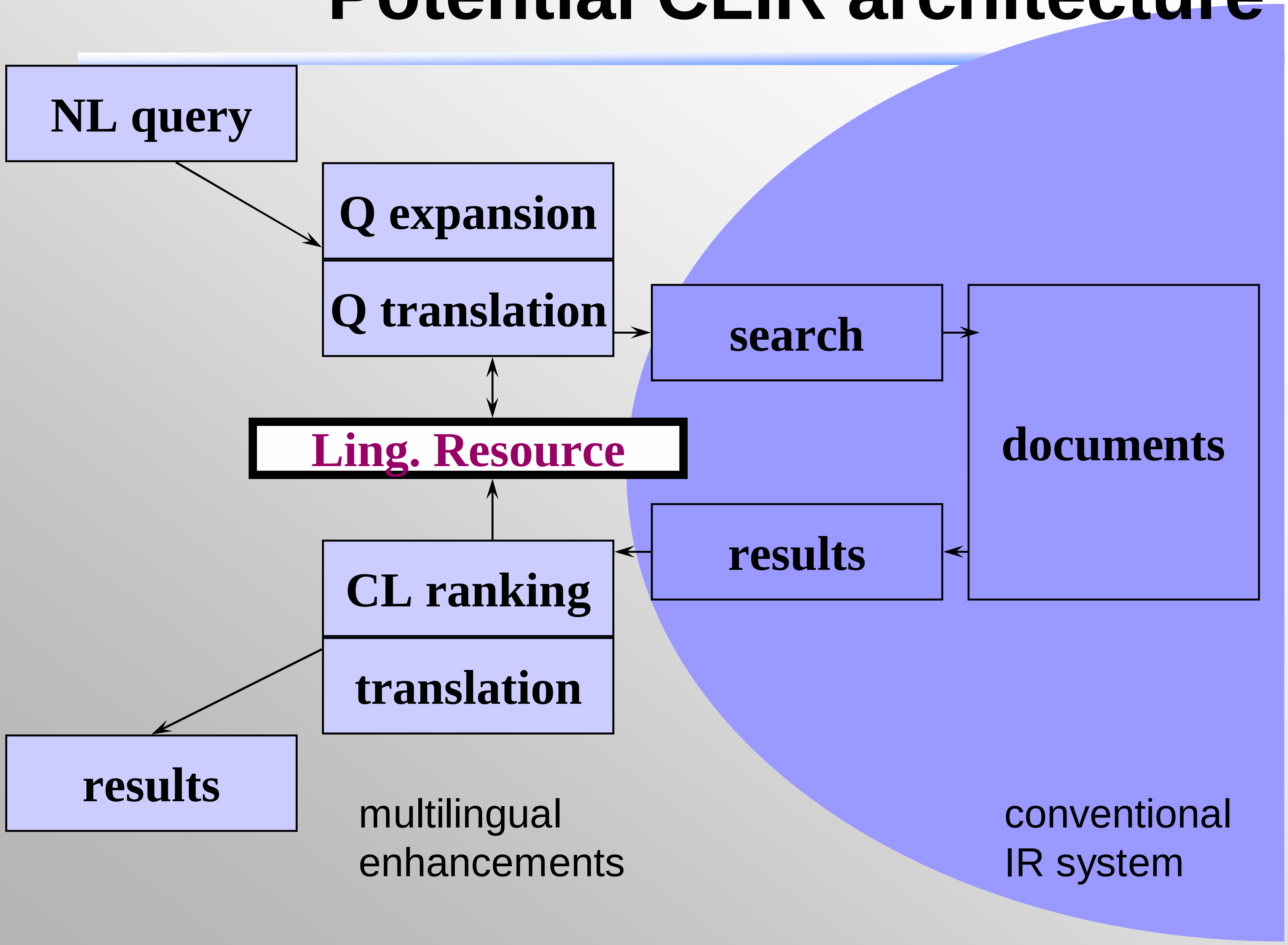
Dok 2:
Europol Edu has confirmed that the number of drugs smuggled in via scheduled flights from Columbia is increasing every year.

Dok 3:
An der ostdeutschen Grenze ist es wiederholt zu Schießereien zwischen rivalisierenden Banden im Drogenkrieg gekommen. 2 Personen wurden getötet, mehrere verletzt.

Translation of query 2

- Approach: Translation query
  - Analyse and translate the query terms
  - Search in (monolingual) Backend-System
- Evaluation
  - Backend database stays unchanged
  - Translation changes do not affect document base
  - Cross-lingual component as system frontend
    - contains multilingual linguistic resource
    - Which is also usable for re-translation
    - And can be maintained independently
  - Crosslinguality is tansparent for the users
  - Fine-tuning between frontend and backend required

Potential CLIR architecture



Outline

III CLIR components

- Document Analysis
- Index creation
- Query processing
  - Translation
  - Expansion
- Ranking
- Result re-translation

Document Analysis 1

- **Language Identification**
  - – Needed to load the proper analysis resources:
    - Tokenisers, stop lists, grammars… depend on language
- **Text Classification**
  - – Used as a content filter: Focus on relevant domains
  - – N-grams or terms as features
- **Named Entity** recognition

  Search for *Bush*$_{\textit{-Person}}$, not for ‚*bush*'

  Search for *Verona*$_{\textit{-Place}}$, not for *Verona*$_{\textit{-Person}}$

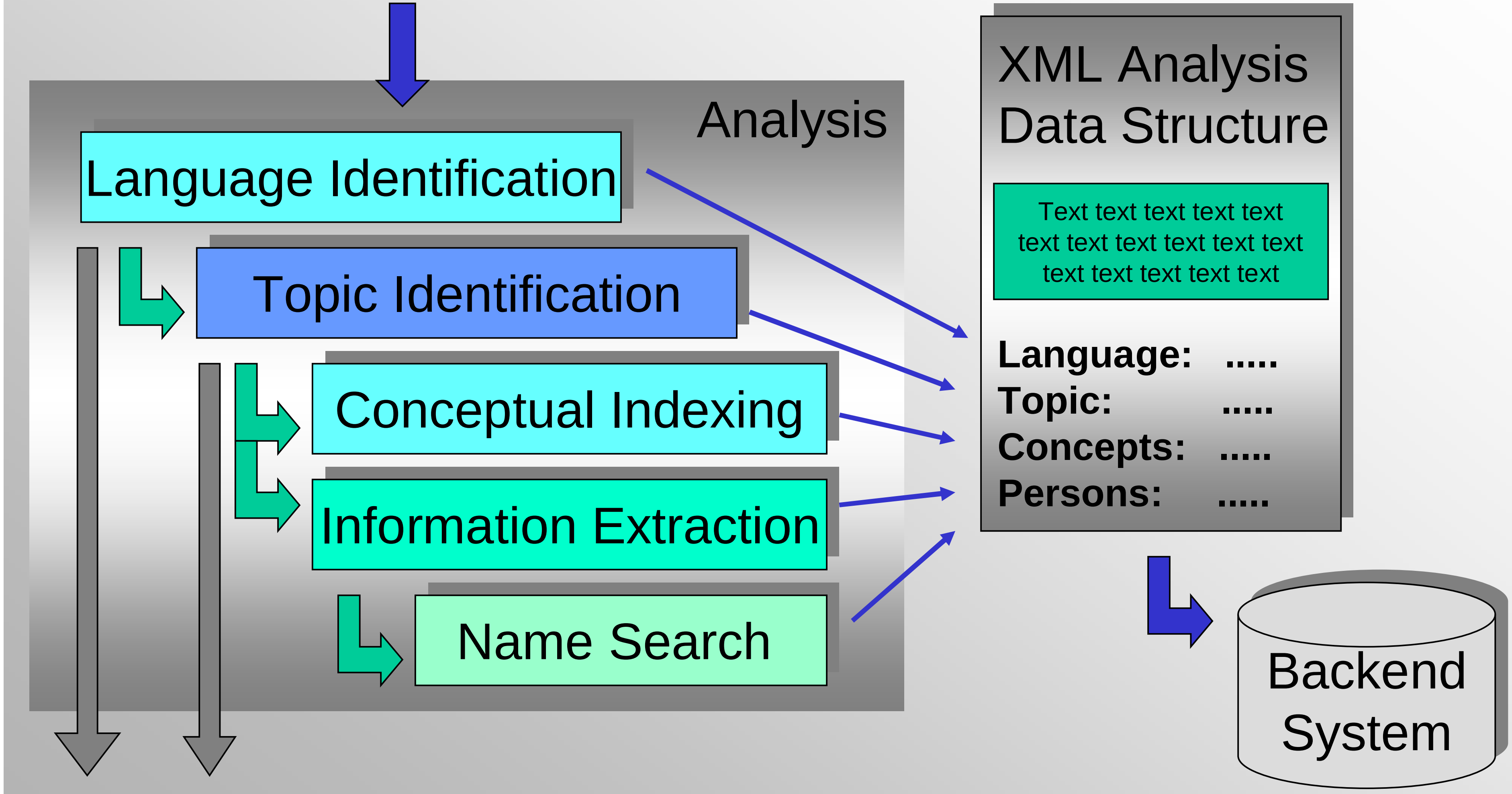
- **Conceptual Indexing**
  - – Find relevant / ‚good' *and translatable* index terms

Document Analysis 2

- Resulting index terms must be **translatable**
  - **Base forms** vs. Text forms vs. Stemming
  - **Concepts** vs. Strings
    - Word senses / Homonyms
    - Concepts are often multiwords:
  - **Multiword** index terms?

- Additional Info: **Normalisation** aspects
  - Capitalisation / Diacritics – foreign characters
  - Mapping of transliteration systems
    - *Yeltsin – Jelzin*, *Abou Moussa – Abbu Mussah*
  - Phonetic search support for names

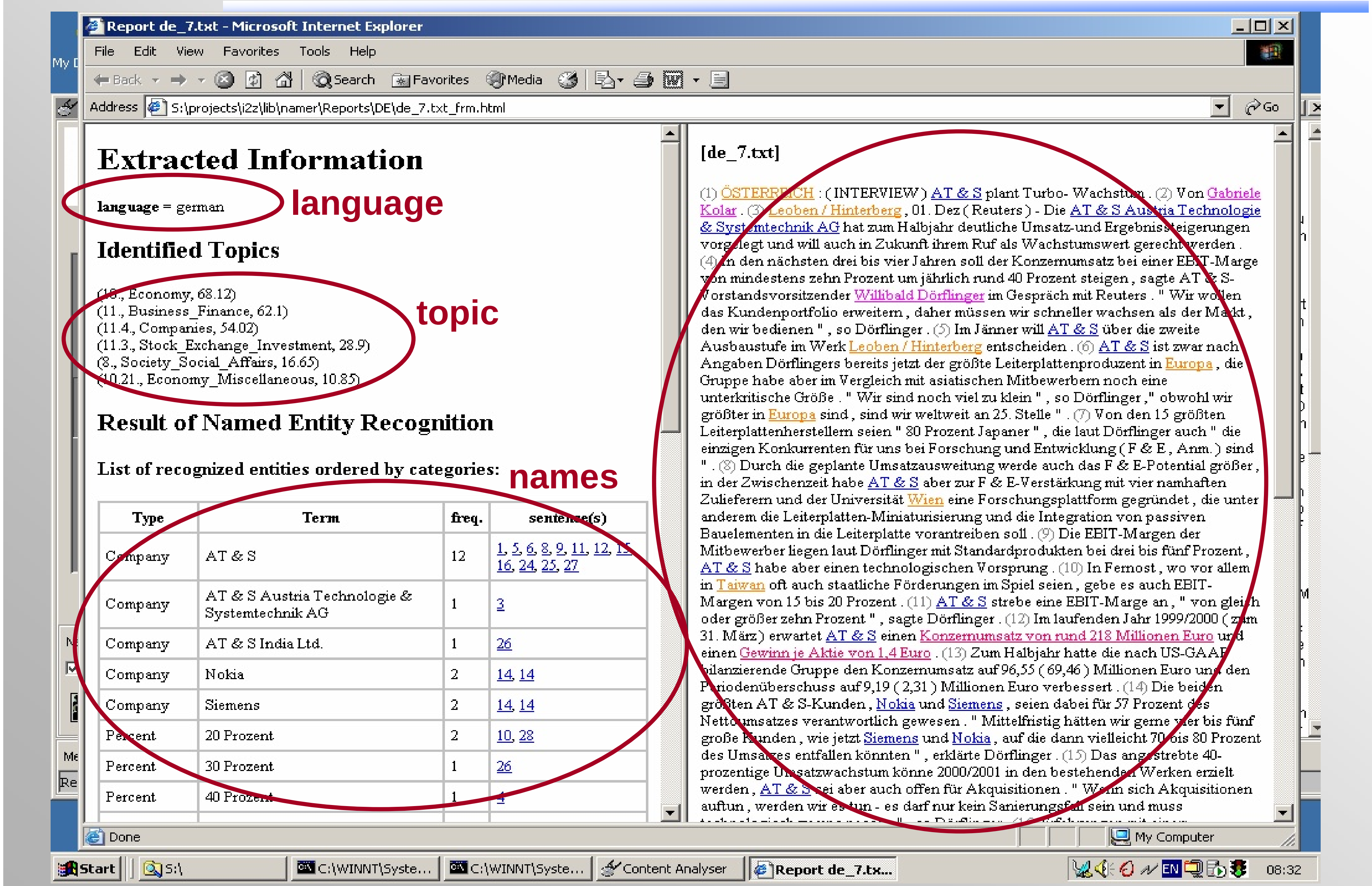
Document Analysis 3



XML Structure for analysis result (backend independent)

Enriching documents with intelligence

Document Analysis 4: Example



Report de_7.txt - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address S:\projects\i2z\lib\namer\Reports\DE\de_7.txt_frm.html
Go
Done
My Computer
Start S:\ C:\WINNT\Syste... C:\WINNT\Syste... Content Analyser Report de_7.tx...
EN 08:32

Extracted Information

language = german

language

Identified Topics

(10., Economy, 68.12)
(11., Business_Finance, 62.1)
(11.4., Companies, 54.02)
(11.3., Stock_Exchange_Investment, 28.9)
(8., Society_Social_Affairs, 16.65)
(10.21., Economy_Miscellaneous, 10.85)

topic

Result of Named Entity Recognition

List of recognized entities ordered by categories:

names

Type	Term	freq.	sentence(s)
Company	AT & S	12	1, 5, 6, 8, 9, 11, 12, 15, 16, 24, 25, 27
Company	AT & S Austria Technologie & Systemtechnik AG	1	3
Company	AT & S India Ltd.	1	26
Company	Nokia	2	14, 14
Company	Siemens	2	14, 14
Percent	20 Prozent	2	10, 28
Percent	30 Prozent	1	26
Percent	40 Prozent	1	4

[de_7.txt]

(1) ÖSTERREICH : (INTERVIEW) AT & S plant Turbo- Wachstum . (2) Von Gabriele Kolar . (3) Leoben / Hinterberg , 01 . Dez (Reuters) - Die AT & S Austria Technologie & Systemtechnik AG hat zum Halbjahr deutliche Umsatz-und Ergebnissteigerungen vorgelegt und will auch in Zukunft ihrem Ruf als Wachstumswert gerecht werden . (4) In den nächsten drei bis vier Jahren soll der Konzernumsatz bei einer EBIT-Marge von mindestens zehn Prozent um jährlich rund 40 Prozent steigen , sagte AT & S-Vorstandsvorsitzender Willibald Dörflinger im Gespräch mit Reuters . " Wir wollen das Kundenportfolio erweitern , daher müssen wir schneller wachsen als der Markt , den wir bedienen " , so Dörflinger . (5) Im Jänner will AT & S über die zweite Ausbaustufe im Werk Leoben / Hinterberg entscheiden . (6) AT & S ist zwar nach Angaben Dörflingers bereits jetzt der größte Leiterplattenproduzent in Europa , die Gruppe habe aber im Vergleich mit asiatischen Mitbewerbern noch eine unterkritische Größe . " Wir sind noch viel zu klein " , so Dörflinger ," obwohl wir größter in Europa sind , sind wir weltweit an 25. Stelle " . (7) Von den 15 größten Leiterplattenherstellern seien " 80 Prozent Japaner " , die laut Dörflinger auch " die einzigen Konkurrenten für uns bei Forschung und Entwicklung (F & E , Anm.) sind " . (8) Durch die geplante Umsatzausweitung werde auch das F & E-Potential größer , in der Zwischenzeit habe AT & S aber zur F & E-Verstärkung mit vier namhaften Zulieferern und der Universität Wien eine Forschungsplattform gegründet , die unter anderem die Leiterplatten-Miniaturisierung und die Integration von passiven Bauelementen in die Leiterplatte vorantreiben soll . (9) Die EBIT-Margen der Mitbewerber liegen laut Dörflinger mit Standardprodukten bei drei bis fünf Prozent , AT & S habe aber einen technologischen Vorsprung . (10) In Fernost , wo vor allem in Taiwan oft auch staatliche Förderungen im Spiel seien , gebe es auch EBIT-Margen von 15 bis 20 Prozent . (11) AT & S strebe eine EBIT-Marge an , " von gleich oder größer zehn Prozent " , sagte Dörflinger . (12) Im laufenden Jahr 1999/2000 (zum 31 . März) erwartet AT & S einen Konzernumsatz von rund 218 Millionen Euro und einen Gewinn je Aktie von 1,4 Euro . (13) Zum Halbjahr hatte die nach US-GAAP bilanzierende Gruppe den Konzernumsatz auf 96,55 (69,46) Millionen Euro und den Periodenüberschuss auf 9,19 (2,31) Millionen Euro verbessert . (14) Die beiden größten AT & S-Kunden , Nokia und Siemens , seien dabei für 57 Prozent des Nettoumsatzes verantwortlich gewesen . " Mittelfristig hätten wir gerne vier bis fünf große Kunden , wie jetzt Siemens und Nokia , auf die dann vielleicht 70 bis 80 Prozent des Umsatzes entfallen könnten " , erklärte Dörflinger . (15) Das angestrebte 40-prozentige Umsatzwachstum könne 2000/2001 in den bestehenden Werken erzielt werden , AT & S sei aber auch offen für Akquisitionen . " Wenn sich Akquisitionen auftun , werden wir es tun - es darf nur kein Sanierungsfall sein und muss

Index organisation 1

- Case 1: one document pool *per language*
  - – Needs a Language Identifier at indexing time
  - – monolingual index per Pool (de, es, fr, ...)
  - – Multiple queries: *one* query *per pool* (de -> de, en -> en, ...)
- Case 2: a *common* document pool
  - – => multilingual index!
  - – Problems with interferences between languages
  - – Problems with Stop lists
  - – One query: *all* queries go to one pool
    - de -> deenfr, en -> deenfr, ...
    - (Language Identifier at runtime)

de-en:	*kind post* *not as*
de-fr:	*Seine ...*
en-fr:	*car ...*

Index organisation 2

- Preselection of documents into monolanguage pools
  - Better Indexing
    - Both for conventional (Stoplisten) and linguistic indexing
  - Simpler search
    - Better Performance: less processing at runtime
    - No inter-language clashes
  - But: Multiple search required

- Option depends on Backend System capabilities
  - Possibility of multiple data pools
  - Language as document attribute
  - …

Query Analysis

Before it can be translated the query must be analysed!

- Find key concepts and translatable units

  „drug and heroin dealers in the former German Democratic Republic“

  – Resolve gapping (*drug dealer / heroin dealer*)

  – Remove untrabslatable (function) words after analysis

  – Identify multiword units

    - Drug dealer / heroin dealer / German Democratic Republic
    - Ingle wod analysis fails …

  – Remove noise words (*former*) (?)

  => Precise linguistic means are required

Query Translation

- Success of CLIR depends of the **quality of translation**:

  Nuclear -> zentral AND power -> Macht AND plant -> Pflanze ☹
  Kernkraftwerk ☺

- **Specialised resource** for the relevant domain
  - No general purpose dictionary helps
  - Even domain specific glossary can be suboptimal

    Glossary has *Droge -> drug* but text has *narcotics*
  - Resource must match the document base

- Problem of **unknown terms**
  - No hits if unknown parts are left in the query, except
  - Proper Names: translate / transliterate …
    - *Persia -> Persien*   (must fit the indexing strategy)

- Problem of target language multiwords (flexer)

Options for Query Translation

Approaches

- Translation by **parallel corpora**
- Translation by **MT**
- Translation using bi-/multilingual **dictionaries**
- Translation using **concept hierarchies**

Problem

- Terminology must match the texts!!

  (Aventinus: coverage ~ 16%, TQPro: coverage ~ 30%)

- Missing translations

Termbank has: *drug -> Medizin*
Texts have : *-> Drogen*

search result: lausy

Query translation by corpora

- Procedure
  - provide parallel corpora (Bible …)
  - Alignment on sentence level
  - Identify corresponding terms (bilingual term alignment)
    - (improvement: multiword terms, lemmatisers)
  - Use those terms for search
- Evaluation
  - Translations match the texts well
  - but: parallel corpora often not available
    - (if so: MLIR possible)
    - New data: gaps in translation

Query translation with MT

- Problems with MT translations
  - short Queries have **few context** for disambiguation
    - Main tproblem: multiword terms
  - MT-dictionaries must fit the **domain**
    - Usually not the case: Babelfish etc. with focus on GV
    - *several* MT-Systems do‘t help (Eurospider Trec8)
    - But: tuned MT came out well in MUMIS (medical IR)
  - 1:n translations must be **disambiguated**
    - How? (By subject area tag?)
  - MT-Systems don‘t have **Ontologies** for Q-Expansion

=> (better way: Generate MT-dictionaries out of CLIR-Resource)

Query Translation using dictionaries

- Ocnventional MRD‘s
  - Look up the Article, Search with all Words?

    much Noise (Univ. New Mexico)
  - Select the reading: Search with „best“ reading?

    often incorrect (Twenty-One in TREC)
  - Formal use often difficult
    - Intended for human lookup
    - Much implicit information (examples …

=> Termbanks / specialised resources

- Selection by using subject area tags
- Expansion by using specialised links (PAROLE / SIMPLE / **NORNA**)

Example: Dictionary Article (LKG)

Urteil

n **1.** judg(e)ment, *der Geschworenen*: verdict (*a.* fig.) (*Straf~, Strafmaß*) sentence (*Scheidungs~*) decree: **das ~ verkünden** pronounce judg(e)ment (*od*. sentence); -> **fällen 2.** (*Ansicht*) judg(e)ment, opinion: **sich ein ~ bilden** form a judg(e)ment (**über** *acc*. about, on); **darüber kann ich mir kein ~ erlauben!** I am no judge (of that)!

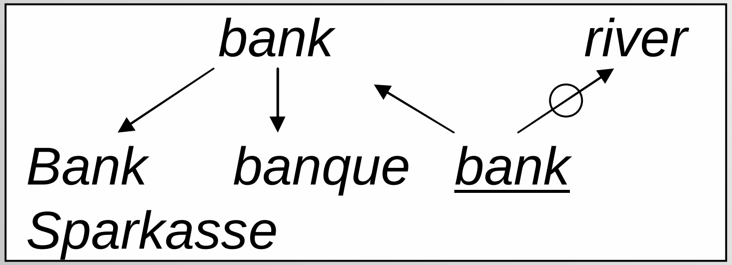
Ton

m **1.** tone, sound (*a. Film, TV, Ggs. Bild*); F **er hat keinen ~ gesagt** he didn‘t say a word; **2.** (*Redeweise*) tone: **ich verbiete mir diesen ~!** I don‘t take that tone with me!; **3.** *ling.* accent, stress: **den ~ legen auf** (acc) *a.fig.* emphasize **4.** (*Farb~*) tone (*a. phot*.), shade: **ein rötlicher ~** a tinge of red.

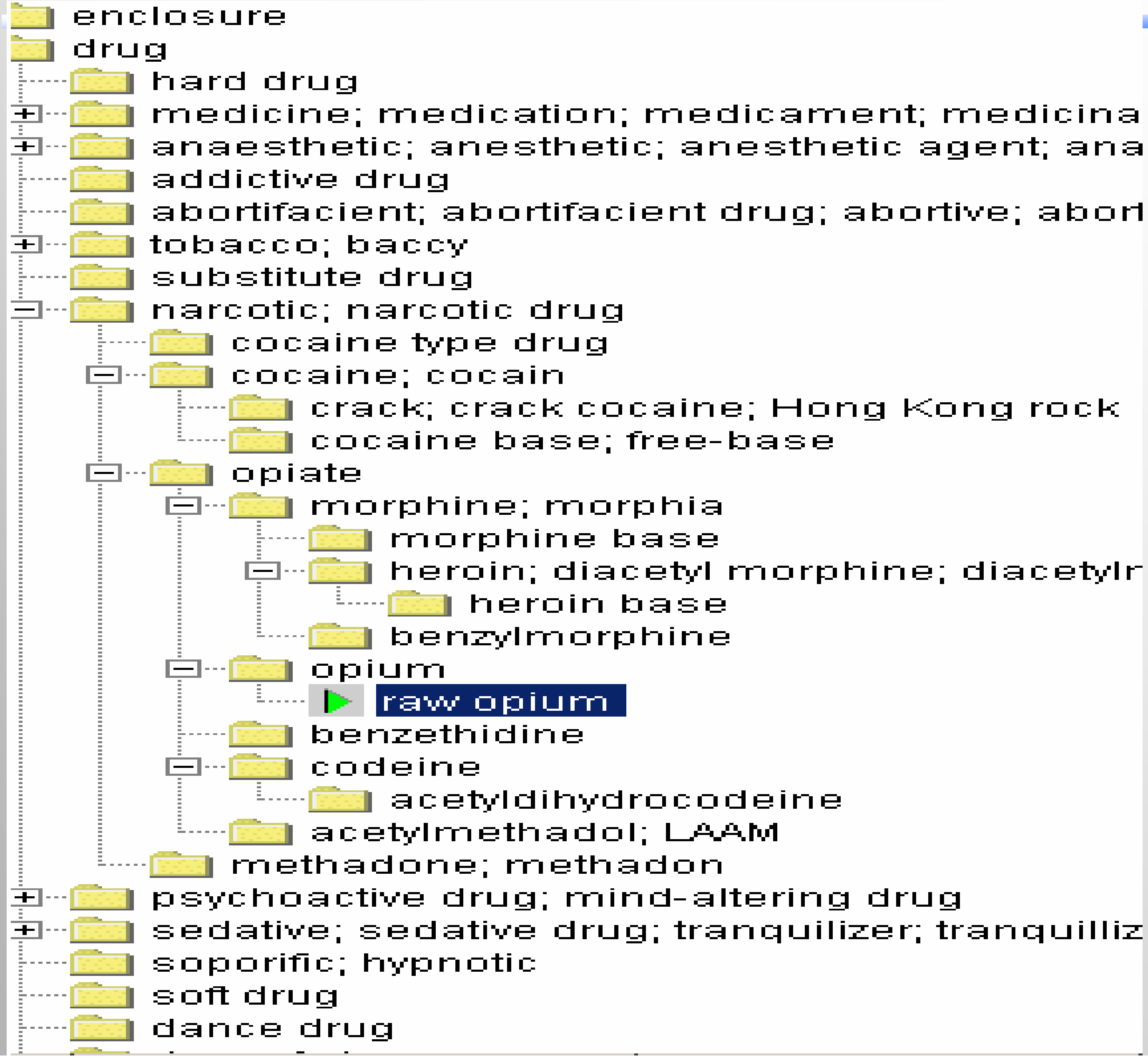
- Univ NewMex takes all targets
- Twenty-One takes first: *judg(e)ment* (0 hits) und *tone*

Translation with concept nets

- Technique:
  - Vreate a conceptional hierarchy („ontology“) where nodes combine clusters of monolingual entries
  - Look up concept for term
  - Look up target terms for concept
- Advantage
  - Integration of Expansion and translation
- Problem
  - Find right concept link (reading)
    - word sense disambiguation: difficult for short Queries
    - Subject area? Often not available, e.g. WordNet



Example concept net

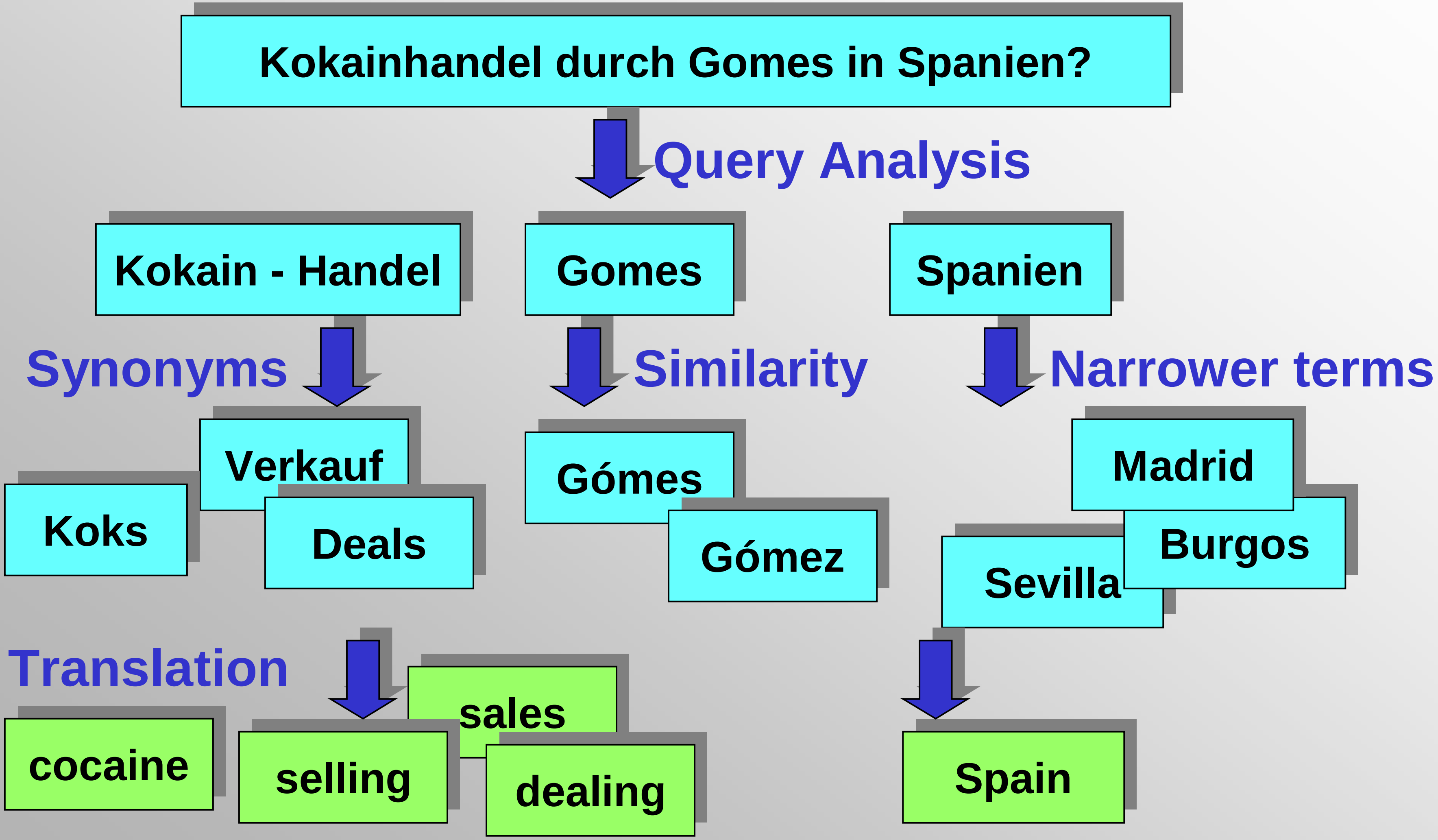


- enclosure
- drug
  - hard drug
  - medicine; medication; medicament; medicina
  - anaesthetic; anesthetic; anesthetic agent; ana
  - addictive drug
  - abortifacient; abortifacient drug; abortive; abort
  - tobacco; baccy
  - substitute drug
  - narcotic; narcotic drug
    - cocaine type drug
    - cocaine; cocain
      - crack; crack cocaine; Hong Kong rock
      - cocaine base; free-base
    - opiate
      - morphine; morphia
        - morphine base
        - heroin; diacetyl morphine; diacetylr
          - heroin base
        - benzylmorphine
      - opium
        - raw opium
      - benzethidine
      - codeine
        - acetyldihydrocodeine
      - acetylmethadol; LAAM
    - methadone; methadon
  - psychoactive drug; mind-altering drug
  - sedative; sedative drug; tranquilizer; tranquilliz
  - soporific; hypnotic
  - soft drug
  - dance drug

Translation: Disambiguation Strategies

- „natural Selection“
  - *plant -> Pflanze / Werk*,
    power plant -> Machtpflanze in no text
- Subject area selection
  - typical MT-Technology: „take terms from IT“
- Specialisation of the linguistic resource
  - WordNet: activate only a segment (IT),
    eliminate potentially wrong readings
- Corpus-Selection
  - bilingual Reference Corpus
  - Take translation with is in the best fitting cluster

Example of QTE



Kokainhandel durch Gomes in Spanien?

Query Analysis

Kokain - Handel

Gomes

Spanien

Synonyms

Similarity

Narrower terms

Verkauf

Koks

Deals

Gómes

Gómez

Madrid

Burgos

Sevilla

Translation

sales

cocaine

selling

dealing

Spain

Query Expansion

- **Conceptual expansion**
  - Searcher‘s and writer‘s formulations need not match
  - Query Expansion: enrich the query by useful alternatives

    Synonyms, narrower or similar terms:
    vehicle? -> car, truck, BMW, …
  - A multilingual concept net helps to do this
    - But: Which concept to pick in case of homonyms

- Expansion on the **document language** side

  If source language synonyms have the same translation, expansion is useless

Persia -> Persien		Persia ->	Persien
Iran -> Persien	☹		SYN Iran ☺

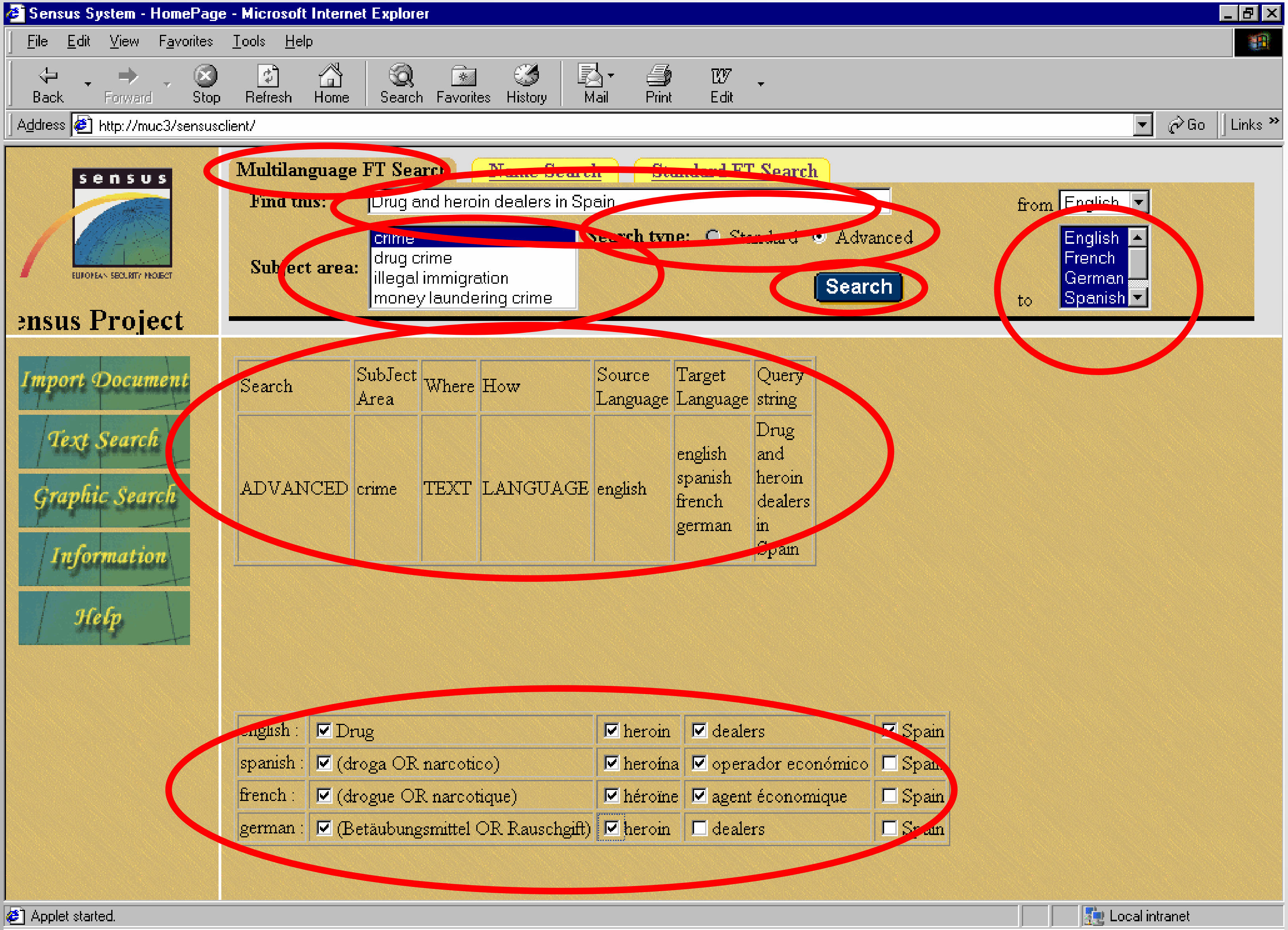
- **Morphological expansion**

  If index has full forms: create all full forms

Expansion techniques

- Document / corpus based
- Concept hierarchy based
  - How deep to go?
- Interactive?
- Expansion is language-specific!

Example: CL Search in Sensus


Sensus System - HomePage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address
Go
Links
sensus
ensus Project
Import Document
Text Search
Graphic Search
Information
Help
Multilanguage FT Search
Name Search
Standard FT Search
Search

Address: http://muc3/sensusclient/

Find this: Drug and heroin dealers in Spain

from English

Search type: ○ Standard ● Advanced

Subject area: crime
drug crime
illegal immigration
money laundering crime

to English
French
German
Spanish

Search	SubJect Area	Where	How	Source Language	Target Language	Query string
ADVANCED	crime	TEXT	LANGUAGE	english	english spanish french german	Drug and heroin dealers in Spain

english :	☑ Drug	☑ heroin	☑ dealers	☑ Spain
spanish :	☑ (droga OR narcotico)	☑ heroína	☑ operador económico	□ Spain
french :	☑ (drogue OR narcotique)	☑ héroïne	☑ agent économique	□ Spain
german :	☑ (Betäubungsmittel OR Rauschgift)	☑ heroin	□ dealers	□ Spain

Applet started.

Local intranet

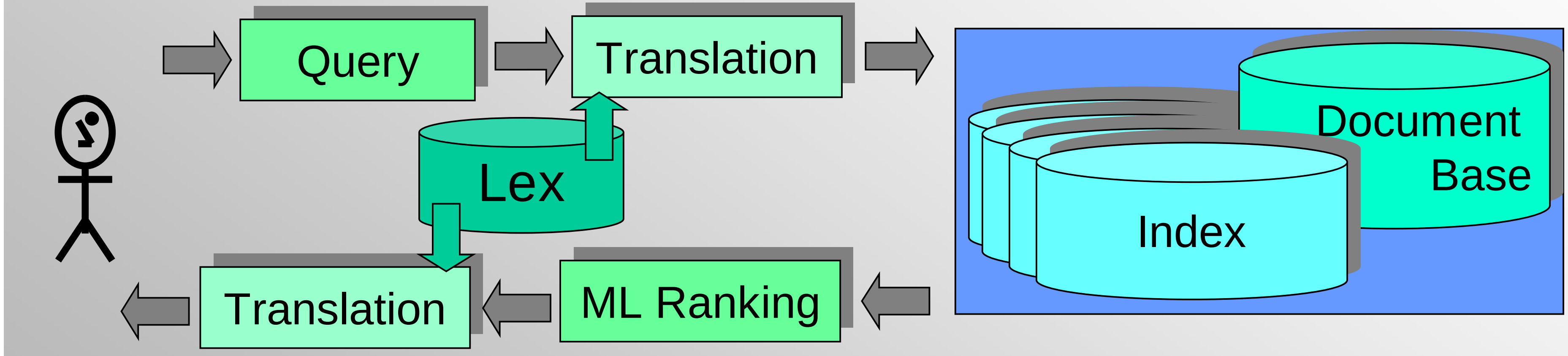
Search and ranking

- Search requires **multiple queries**
  - Select proper index per language
    - English query to Englich index
      German query to German index, …
  - Do textbase/language-specific (statistical) expansion
  - Come back with several document sets
    - E.g.: 30 hits in French, 12 in German, 25 in English

- **Ranking**
  - Is a hit in French better than a hit in English?
  - Current ranking approaches are monolingual
    - Based on a common set of terms (found/nonfound):
      What if the term sets are completely disjoint?
  - Alternatives:
    - Rank the languages internally, present them separated
    - Find a common basis of comparison (e.g. translation)
    - Create a sequence of queries with released constraints

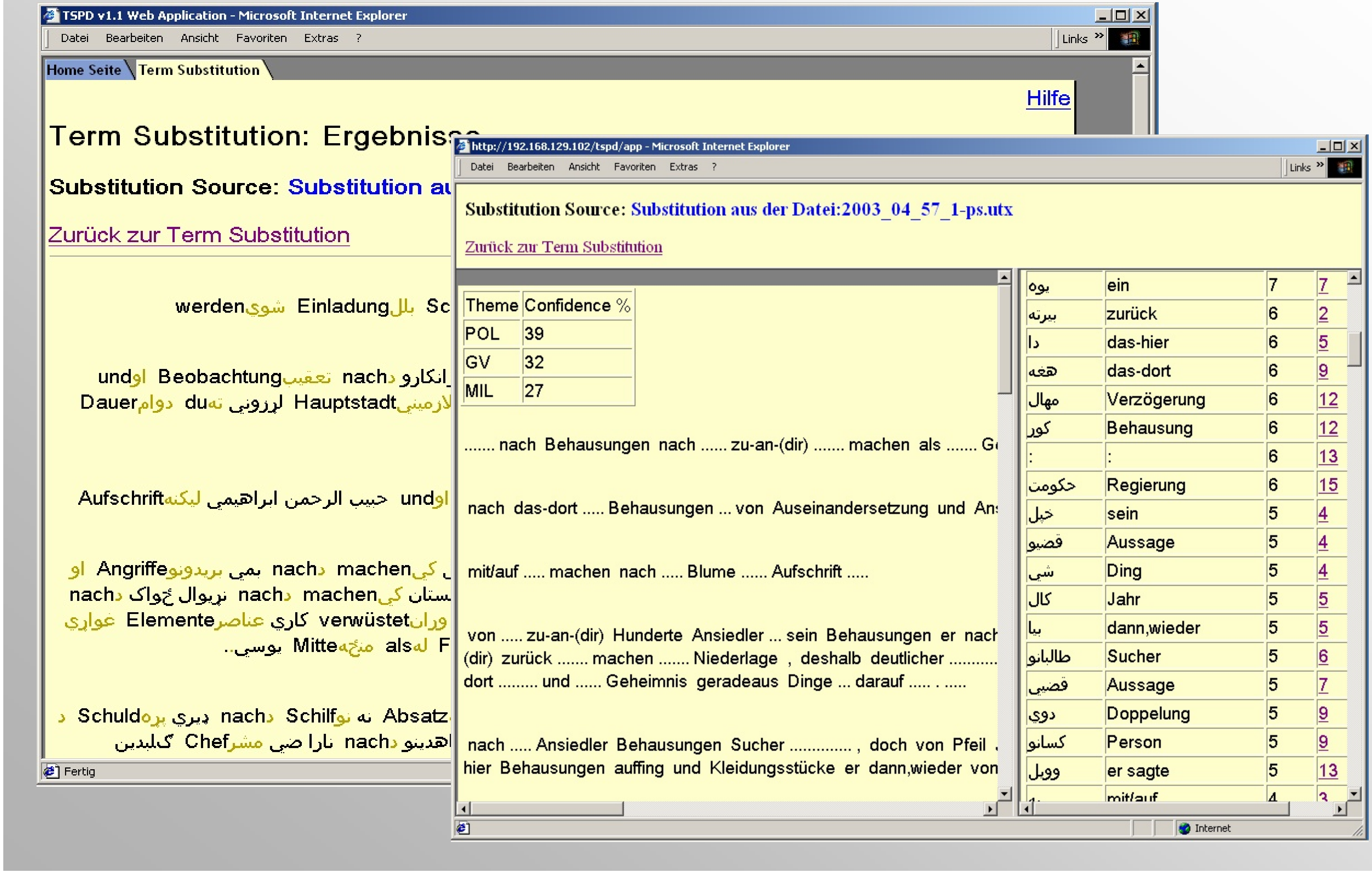
Result translation

- Foreign language search results must be brought back into the interface language! (NORNA)



- Options:
  - Machine translation
  - Term replacement
- Target: understand the content / judge relevancy

Example: Term Substitution Output


TSPD v1.1 Web Application - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Links
Home Seite
Term Substitution
Hilfe
Fertig
http://192.168.129.102/tspd/app - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Links
Internet

Term Substitution: Ergebnisse

Substitution Source: Substitution au

Zurück zur Term Substitution

werdenشوي Einladungبلل Sc

undاو Beobachtungتعقیب nachد انکارو
Dauerدوام duته لړزوني Hauptstadtلازميني

Aufschriftلیکنه حبیب الرحمن ابراهیمي undاو

او Angriffeبریدونو بمي nachد machenکي
nachد ځواک نړیوال nachد machenکي ستان
غواړي Elementeعناصر کاري verwüstetوران
..یوسي Mitteمنځه alsله F

د Schuldپړه ډېري nachد Schilfنو نه Absatz
ګلبدین Chefمشر نارا ضي nachد اهدینو

Substitution Source: Substitution aus der Datei:2003_04_57_1-ps.utx

Zurück zur Term Substitution

Theme	Confidence %
POL	39
GV	32
MIL	27

.... nach Behausungen nach ...... zu-an-(dir) .... machen als .... G

nach das-dort ..... Behausungen ... von Auseinandersetzung und An

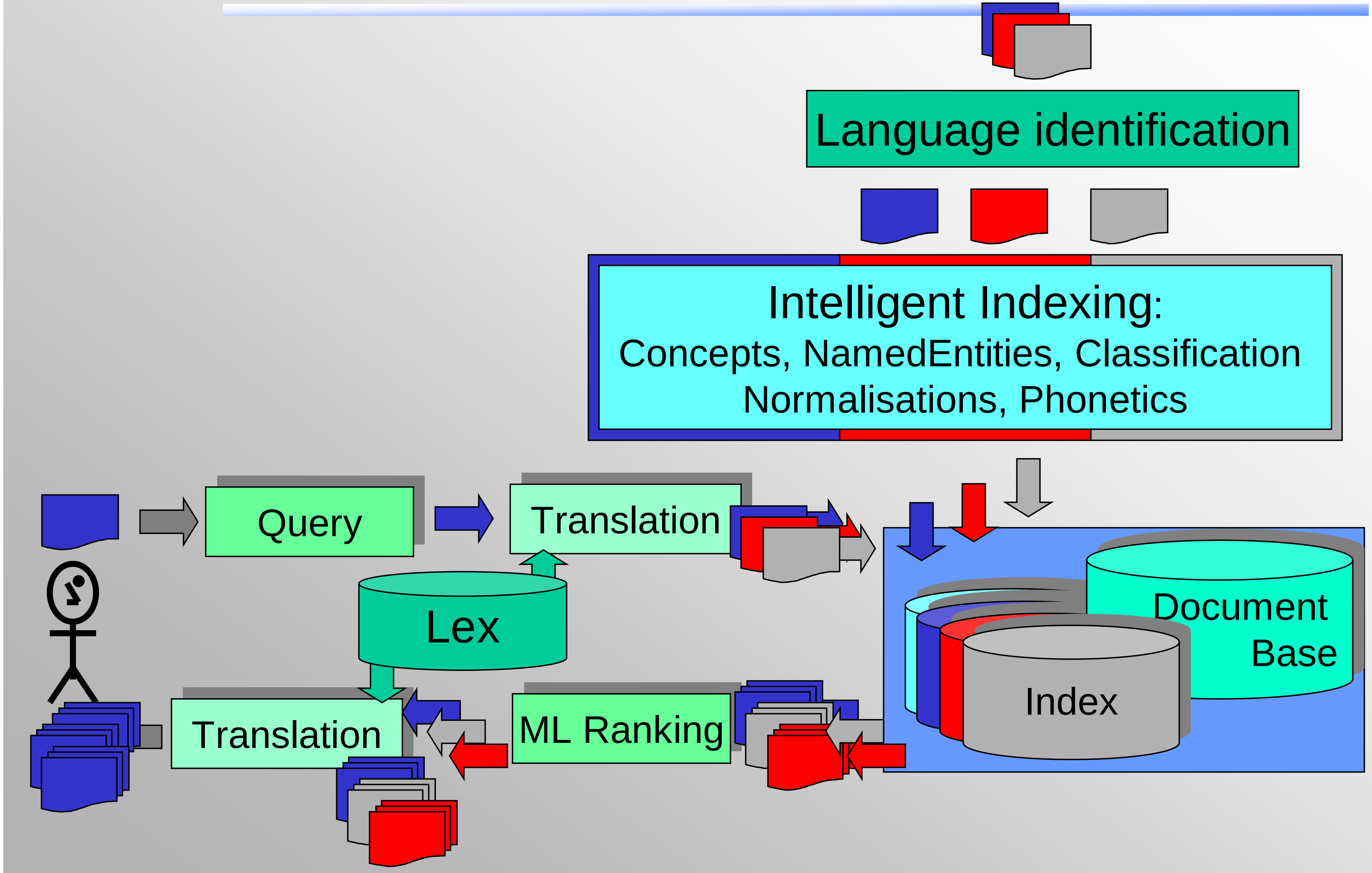
mit/auf ..... machen nach ..... Blume ...... Aufschrift .....

von ..... zu-an-(dir) Hunderte Ansiedler ... sein Behausungen er nach
(dir) zurück .... machen .... Niederlage , deshalb deutlicher ....
dort .... und ...... Geheimnis geradeaus Dinge ... darauf ....

nach ..... Ansiedler Behausungen Sucher .... , doch von Pfeil .
hier Behausungen auffing und Kleidungsstücke er dann,wieder von

یوه	ein	7	7
بیرته	zurück	6	2
دا	das-hier	6	5
هغه	das-dort	6	9
مهال	Verzögerung	6	12
کور	Behausung	6	12
:	:	6	13
حکومت	Regierung	6	15
خپل	sein	5	4
قضیو	Aussage	5	4
شي	Ding	5	4
کال	Jahr	5	5
بیا	dann,wieder	5	5
طالبانو	Sucher	5	6
قضیې	Aussage	5	7
دوي	Doppelung	5	9
کسانو	Person	5	9
وویل	er sagte	5	13
	mit/auf	4	3

Complete System



Consequence for the resource

Engineering aspects when implementing a multilingual information processing system

- Focus on cross-lingual information retrieval
- Features of such a system containing multilingual extensions
- Requirements for **resources** to support such a system
- Maintenance and build-up of such resources

Resources for Crosslingual Retrieval

CLIR are the most complex ones for resources!

1. **Document Analysis** requires linguistic analysis (TREC)
   - lemmatiser, head-modifier-parser, word sense disambiguation
2. **Query Expansion** + Feedback improves search results significantly (TREC)
3. **Crosslinguality** requires translation of queries and hit documents (CLEF)

All these steps need **linguistic resources**

Resources must be **maintained**

- Tools to build them
- Tools for maintenance: Single point of administration
- Tools for exchange: Easy exchange

Resources for Document Analysis

Monolingual lexical resource, covering

Minimum: **Morphosyntax** for index term creation

- POS and inflection
- (additional requirement for CLIR: multiword terms!)

+: information on **argument structures** etc.

- Dependency trees / head-modifier analysis

+: Some **semantic and collocation** information

- Named Entity recognition
- Word sense disambiguation

Resources for Query Expansion

- Expansion needs **relations** of a **ConceptNet**
  - Strict hierarchical relations (automatic expansion)
    - Synonyms, narrower terms
  - Additional relations (interactive expansion)
    - Part-of, see-also, ...

- ConceptNet must be **descriptive**
  - Contains all searchable terms of a corpus
    - Unlike thesaurus (which is normative)

  => To be built for each single application
    - ConceptNet Building Tools are important!

Query and Document Translation

- Query Translation needs **terminology data**
  - Concept-oriented term banks
  - 50-70% multiword terms
- Document Retranslation needs **MT lexicons**
  - Monolingual and bilingual (transfer) information
  - Reading vs. Concept organisation
  - Transfers: tests&actions, statistical clusters
- Translation resources must be **consistent**

  => Exchange requirements between system components
- Translation resources must **fit the text base**

  => Building tools

Resources: Requirements

CLIR needs a resource like the following:

- Organised as a **Multilingual ConceptNet**
  - (language-specific wordnets do not really help in CLIR)
- Nodes: **Concepts**
  - Concepts are abstract, defined by definitions
  - Concepts have **multilingual expressions** („concept“)

    -> **Query translation**
  - Each monolingual concept has its terms
  - Terms have (monolingual) **linguistic descriptions**

    -> **Indexing / Analysis**
- Links: **Relations**
  - Strict hierarchies (is_a), additional relations

    -> **Query expansion**

ConceptNet vs. Dictionary

- ConceptNet is not a dictionary
  - **Organisation**
    - Dictionary is organised in **lemmata**
      - All concepts with same lemma form a dictionary *article*
    - ConceptNet is organised in **concepts**
      - Lemmas for different concepts are different entries
  - **Relations**
    - Dictionary does not explicit hierarchical relations
    - Concepts use relations to build **hierarchies** & networks
- But: It uses dictionary information
  - (basic linguistic information, subject area)

ConceptNet vs. Thesaurus

- ConceptNet is not a thesaurus / ontology
  - Thesaurus is based on **canonical** terms
    - Thesaurus terms *represent* concepts
      other words are not to be used
  - ConceptNet is based on **used** terms
    - Many terms can represent a concept
    - Analysis base for a ConceptNet is a *corpus*
      all terms in the corpus are in the ConceptNet
- But: Both model **relations** between concepts
  - ConceptNets have more terms (synonyms)
  - ConceptNets have more relation types (EuroWordNet)

ConceptNet vs. Terminology Entry

- Terminology entry is canonical, not descriptive
- Assumes 1:1 relationship between languages
- Does not provide:
  - Explicit hierarchical structure
  - Detailed linguistic annotations
  - Idiosyncratic relations between languages
    - Esp.: transfer tests and actions
(can differ from en > fr and da > de within one term!)
- But: Term metamodel similar to Concept nodes
  - Cf. TBX standard – not MILE compatible ☺

ConceptNet vs. WordNet

- Similar: based on a kind of SynSet representation
- Definition of concept / synset node

- But:
  - WordNet is not multilingual
    - One hierarchy (per language)
    - EuroWordNet – ILI: ?
  - No linguistc annotations per term
  - No relationships between synset terms
    - (On transfer level: tests & actions / preferences)
    - On crossreference level: abbreviations, headwords, …

- WordNet is not a WordNet

Are ConceptNets domain-independent?

- No. Multilingual ConceptNet is **domain specific**
  - Concept hierarchy is purpose-dependent
  - Users will do searches in their domain
    - Search terms sometimes are user-specific
    - Prepare resource based on user-corpus
  - General purpose terms are difficult to search & translate
    - 1:n translations introduce noise in the search
    - People tend to use specific terms for searching
- Link different ConceptNet instances
  - Use a common **top level ontology**

Are ConceptNets language independent?

- **Yes!**
  - As far as CLIR applications are concerned
    - Language-independent definition of concepts
      - Definition by: defintion text, part-of-speech, subject-field
    - Language-independent definition of relations
      - *Floppy_Drives* are *part_of computers* in any language
- **No!**
  - As far as linguistic accuracy is concerned
    - Concepts are defined by (connotational) context
    - Concept hierarchies are sometimes language-specific
      - (legal system, educational system, ...)

=> Cover concepts in one ConceptNet

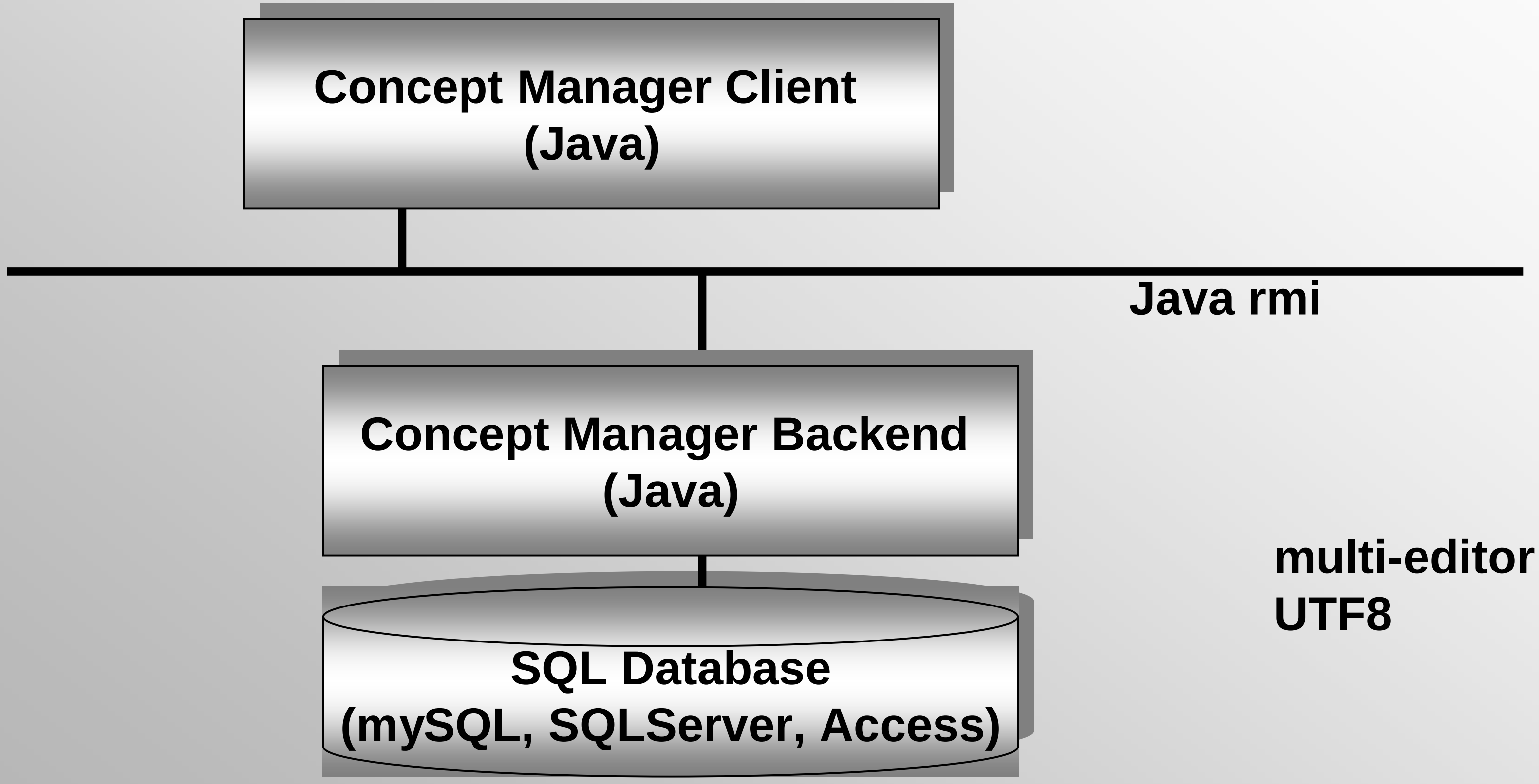
Some concepts do not have terms for some languages

Result: Linguistic Design

- Organise resources in **concepts**
  - Homonyms create noise in retrieval
- Base concepts on an ontology / hierarchical **net**
  - Start from a top level ontology (EuroWordNet)
  - Application specific hierarchy below the top level
  - Allow for additional relations (subset of EuroWordNet)
- Describe concepts with **terms (entries)**
  - Cover all relevant terms in the domain
- Enrich concepts with **multilingual** terms / entries
  - To be used as translations in query and text translation
- Describe terms with linguistic **annotations**
  - Needed for compilation into linguistic applications
  - Based on OLIF / EAGLES / MILE

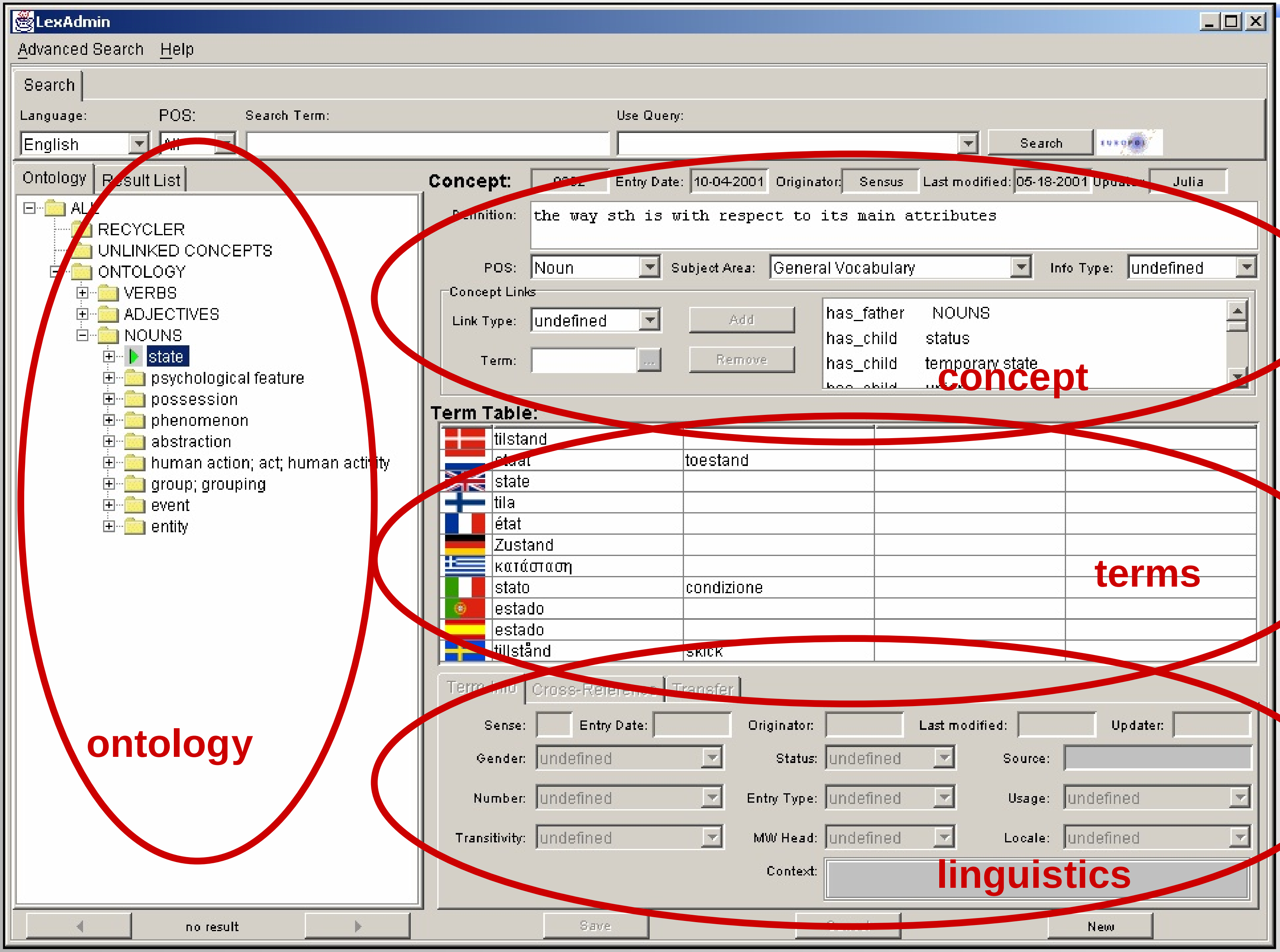
IV ConceptManager: Software Structure

Software Structure



multi-editor
UTF8

ConceptNet Manager: GUI



Language: English
POS: All
Search Term:
Use Query:

- ALL
  - RECYCLER
  - UNLINKED CONCEPTS
  - ONTOLOGY
    - VERBS
    - ADJECTIVES
    - NOUNS
      - state
      - psychological feature
      - possession
      - phenomenon
      - abstraction
      - human action; act; human activity
      - group; grouping
      - event
      - entity

Concept: 0002 Entry Date: 10-04-2001 Originator: Sensus Last modified: 05-18-2001 Updater: Julia

Definition: the way sth is with respect to its main attributes

POS: Noun Subject Area: General Vocabulary Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father	NOUNS
has_child	status
has_child	temporary state

Term Table:

	tilstand			
	staat	toestand		
	state			
	tila			
	état			
	Zustand			
	κατάσταση			
	stato	condizione		
	estado			
	estado			
	tillstånd	skick		

Sense: Entry Date: Originator: Last modified: Updater:

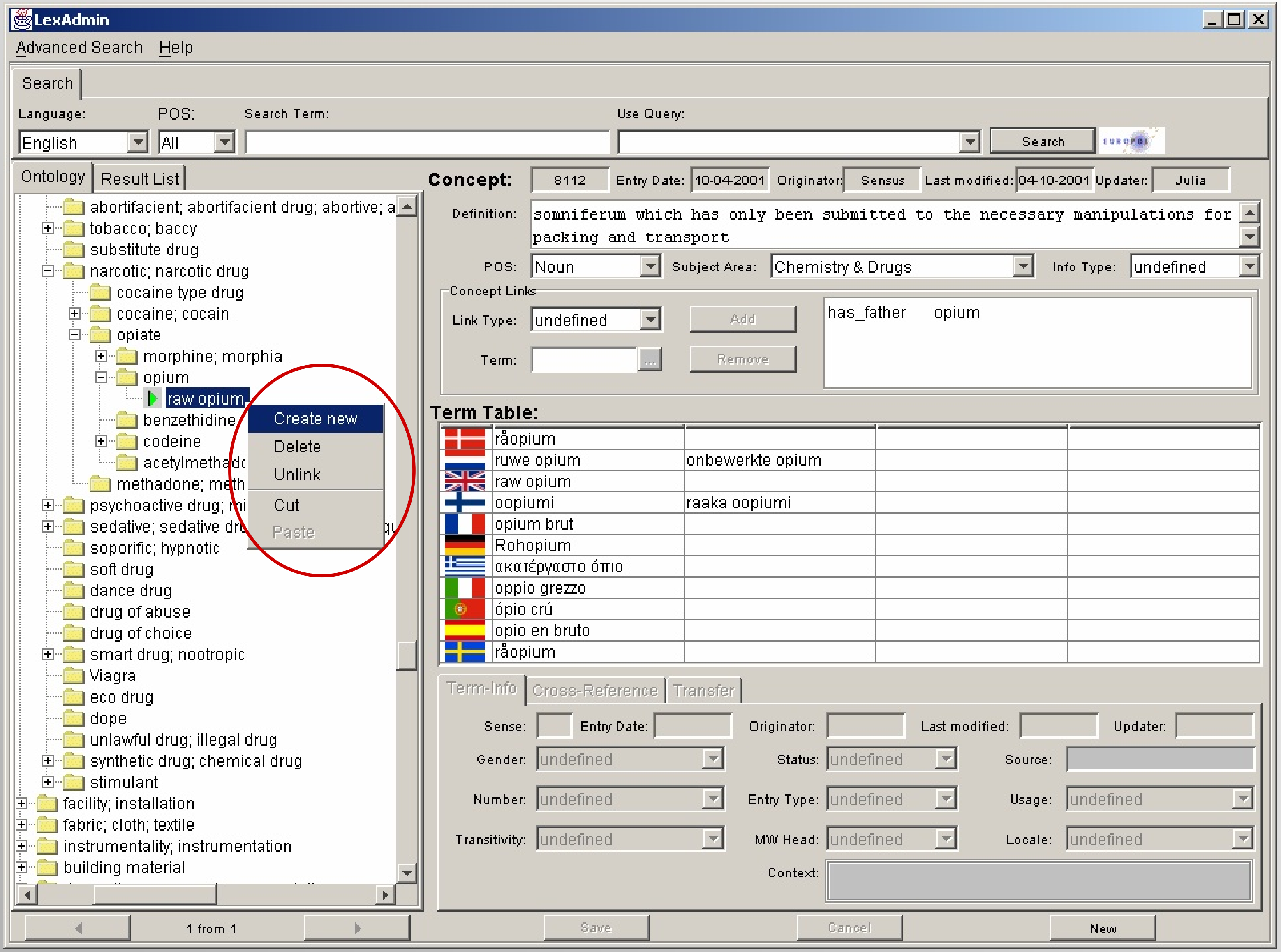
Gender: undefined Status: undefined Source:

Number: undefined Entry Type: undefined Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Ontology Tree: Menu



Language: English POS: All Search Term: Use Query:

- abortifacient; abortifacient drug; abortive; a
- tobacco; baccy
- substitute drug
- narcotic; narcotic drug
  - cocaine type drug
  - cocaine; cocain
  - opiate
    - morphine; morphia
    - opium
      - raw opium
    - benzethidine
    - codeine
    - acetylmethadc
  - methadone; meth
- psychoactive drug; mi
- sedative; sedative dru
- soporific; hypnotic
- soft drug
- dance drug
- drug of abuse
- drug of choice
- smart drug; nootropic
- Viagra
- eco drug
- dope
- unlawful drug; illegal drug
- synthetic drug; chemical drug
- stimulant
- facility; installation
- fabric; cloth; textile
- instrumentality; instrumentation
- building material

Concept: 8112 Entry Date: 10-04-2001 Originator: Sensus Last modified: 04-10-2001 Updater: Julia

Definition: somniferum which has only been submitted to the necessary manipulations for packing and transport

POS: Noun Subject Area: Chemistry & Drugs Info Type: undefined

Concept Links

Link Type: undefined Term:

has_father opium

Term Table:

🇩🇰	råopium			
🇳🇱	ruwe opium	onbewerkte opium		
🇬🇧	raw opium			
🇫🇮	oopiumi	raaka oopiumi		
🇫🇷	opium brut			
🇩🇪	Rohopium			
🇬🇷	ακατέργαστο όπιο			
🇮🇹	oppio grezzo			
🇵🇹	ópio crú			
🇪🇸	opio en bruto			
🇸🇪	råopium			

Sense: Entry Date: Originator: Last modified: Updater:

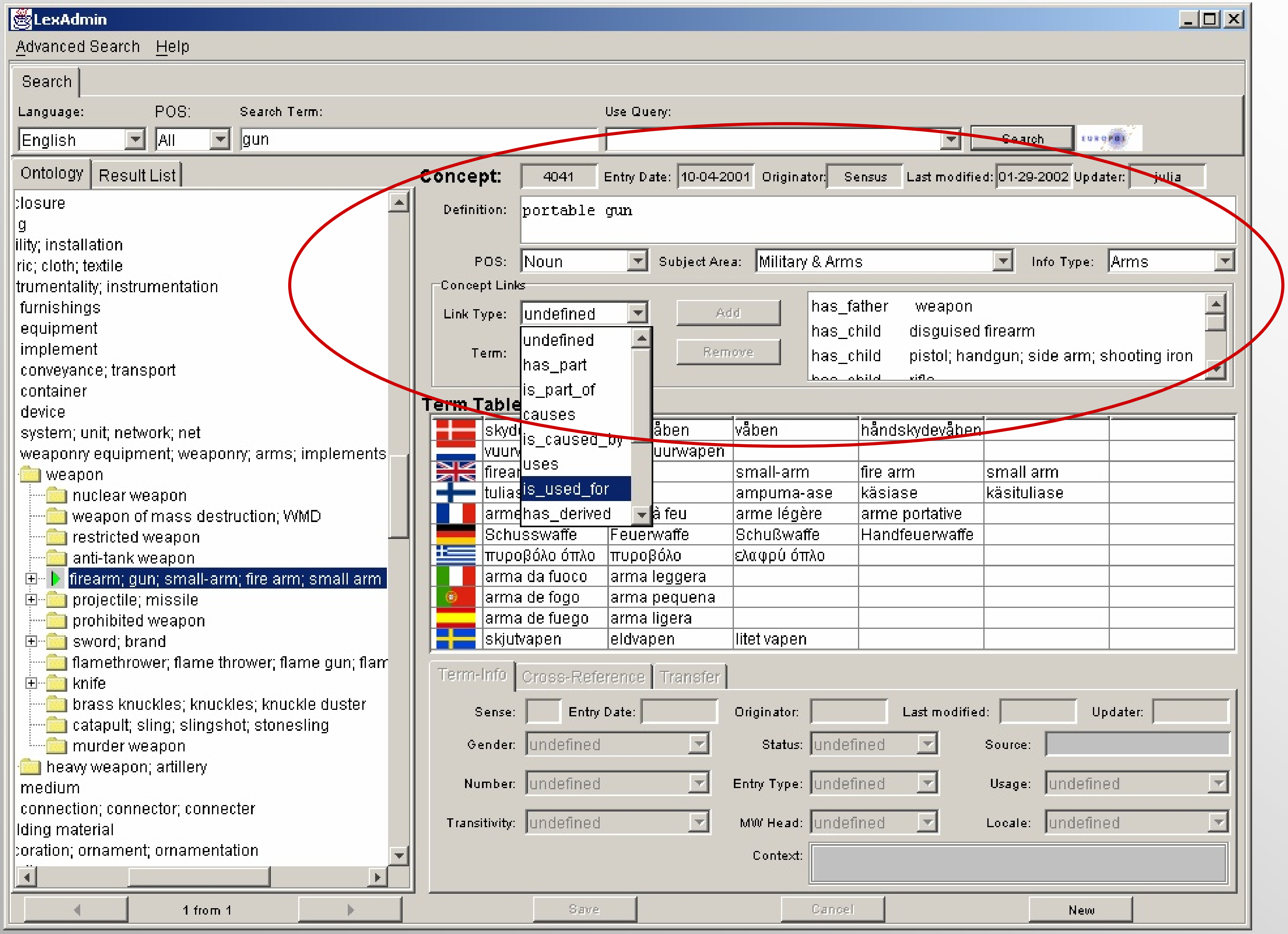
Gender: undefined Status: undefined Source:

Number: undefined Entry Type: undefined Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Active Concept: Concept Features



Language: English

POS: All

Search Term: gun

Use Query:

:losure
g
ility; installation
ric; cloth; textile
trumentality; instrumentation
furnishings
equipment
implement
conveyance; transport
container
device
system; unit; network; net
weaponry equipment; weaponry; arms; implements
weapon
nuclear weapon
weapon of mass destruction; WMD
restricted weapon
anti-tank weapon
firearm; gun; small-arm; fire arm; small arm
projectile; missile
prohibited weapon
sword; brand
flamethrower; flame thrower; flame gun; flam
knife
brass knuckles; knuckles; knuckle duster
catapult; sling; slingshot; stonesling
murder weapon
heavy weapon; artillery
medium
connection; connector; connecter
lding material
:oration; ornament; ornamentation

Concept: 4041 Entry Date: 10-04-2001 Originator: Sensus Last modified: 01-29-2002 Updater: julia

Definition: portable gun

POS: Noun Subject Area: Military & Arms Info Type: Arms

Concept Links

Link Type: undefined

Term:

undefined
has_part
is_part_of
causes
is_caused_by
uses
is_used_for
has_derived

has_father	weapon
has_child	disguised firearm
has_child	pistol; handgun; side arm; shooting iron
has_child	rifle

Term Table

	skyd	åben	våben	håndskydevåben		
	vuurv	uurwapen				
	firear		small-arm	fire arm	small arm	
	tulias		ampuma-ase	käsiase	käsituliase	
	arme	à feu	arme légère	arme portative		
	Schusswaffe	Feuerwaffe	Schußwaffe	Handfeuerwaffe		
	πυροβόλο όπλο	πυροβόλο	ελαφρύ όπλο			
	arma da fuoco	arma leggera				
	arma de fogo	arma pequena				
	arma de fuego	arma ligera				
	skjutvapen	eldvapen	litet vapen			

Sense: Entry Date: Originator: Last modified: Updater:

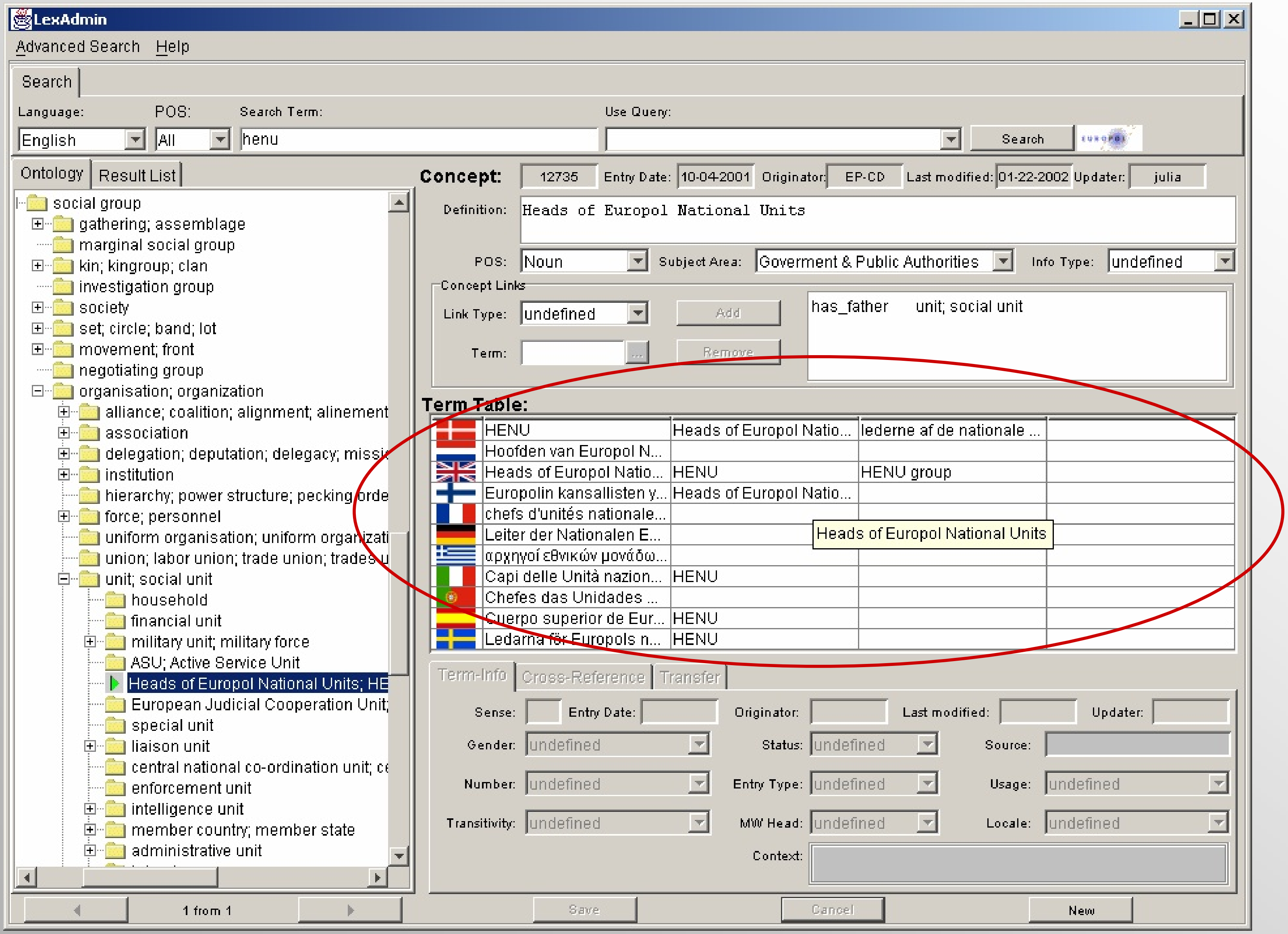
Gender: undefined Status: undefined Source:

Number: undefined Entry Type: undefined Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Active Concept: Term Table



Language: English
POS: All
Search Term: henu
Use Query:

- social group
  - gathering; assemblage
  - marginal social group
  - kin; kingroup; clan
  - investigation group
  - society
  - set; circle; band; lot
  - movement; front
  - negotiating group
  - organisation; organization
    - alliance; coalition; alignment; alinement
    - association
    - delegation; deputation; delegacy; missi
    - institution
    - hierarchy; power structure; pecking orde
    - force; personnel
    - uniform organisation; uniform organizati
    - union; labor union; trade union; trades u
    - unit; social unit
      - household
      - financial unit
      - military unit; military force
      - ASU; Active Service Unit
      - Heads of Europol National Units; HE
      - European Judicial Cooperation Unit;
      - special unit
      - liaison unit
      - central national co-ordination unit; ce
      - enforcement unit
      - intelligence unit
      - member country; member state
      - administrative unit

Concept: 12735 Entry Date: 10-04-2001 Originator: EP-CD Last modified: 01-22-2002 Updater: julia

Definition: Heads of Europol National Units

POS: Noun Subject Area: Goverment & Public Authorities Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father unit; social unit

Term Table:

🇩🇰	HENU	Heads of Europol Natio...	lederne af de nationale ...	
🇳🇱	Hoofden van Europol N...			
🇬🇧	Heads of Europol Natio...	HENU	HENU group	
🇫🇮	Europolin kansallisten y...	Heads of Europol Natio...		
🇫🇷	chefs d'unités nationale...			
🇩🇪	Leiter der Nationalen E...			
🇬🇷	αρχηγοί εθνικών μονάδω...			
🇮🇹	Capi delle Unità nazion...	HENU		
🇵🇹	Chefes das Unidades ...			
🇪🇸	Cuerpo superior de Eur...	HENU		
🇸🇪	Ledarna för Europols n...	HENU		

Heads of Europol National Units

Sense: Entry Date: Originator: Last modified: Updater:

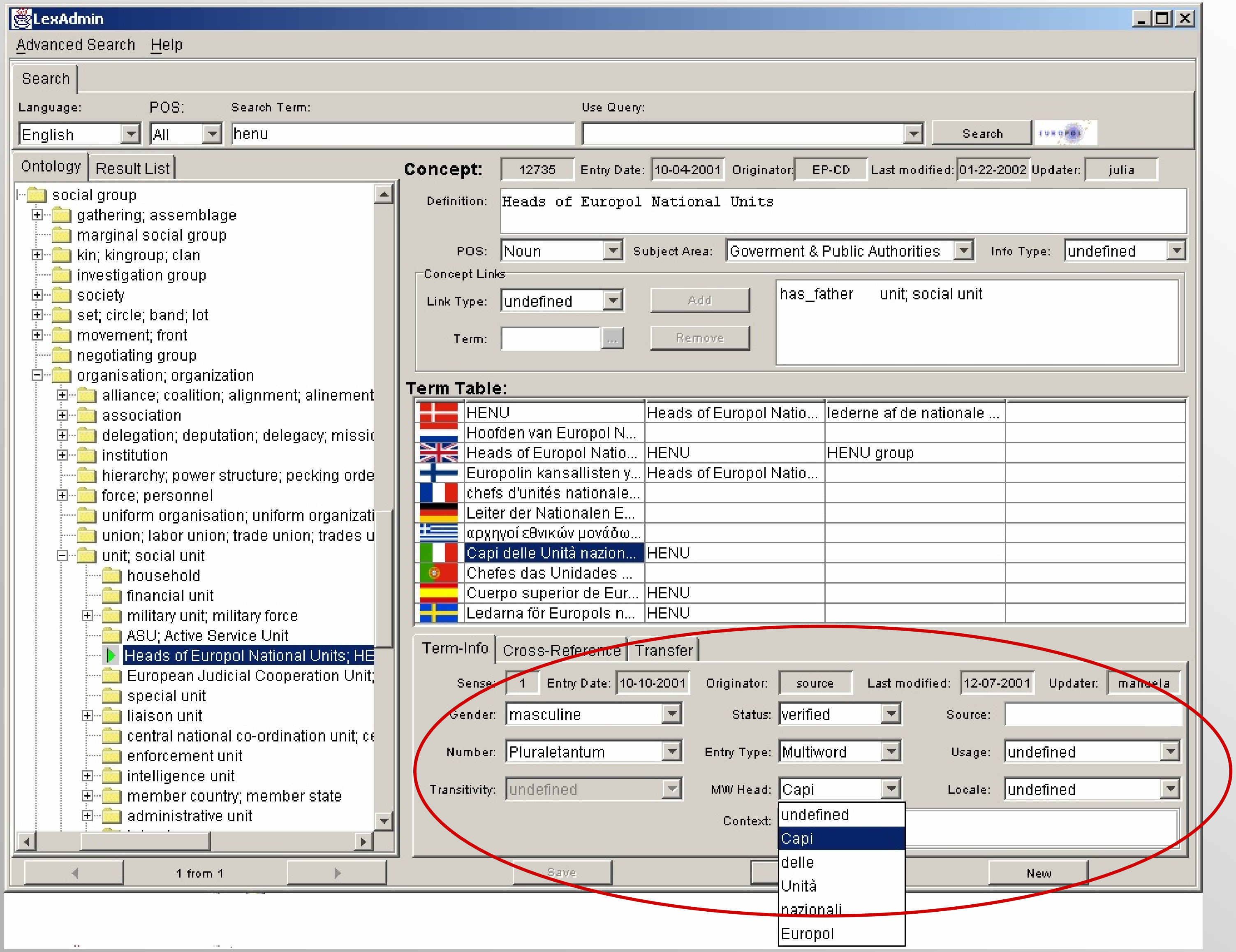
Gender: undefined Status: undefined Source:

Number: undefined Entry Type: undefined Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Term Table: Linguistic Info



Language: English
POS: All
Search Term: henu
Use Query:

- social group
  - gathering; assemblage
  - marginal social group
  - kin; kingroup; clan
  - investigation group
  - society
  - set; circle; band; lot
  - movement; front
  - negotiating group
  - organisation; organization
    - alliance; coalition; alignment; alinement
    - association
    - delegation; deputation; delegacy; missic
    - institution
    - hierarchy; power structure; pecking orde
    - force; personnel
    - uniform organisation; uniform organizati
    - union; labor union; trade union; trades u
    - unit; social unit
      - household
      - financial unit
      - military unit; military force
      - ASU; Active Service Unit
      - Heads of Europol National Units; HE
      - European Judicial Cooperation Unit;
      - special unit
      - liaison unit
      - central national co-ordination unit; ce
      - enforcement unit
      - intelligence unit
      - member country; member state
      - administrative unit

Concept: 12735 Entry Date: 10-04-2001 Originator: EP-CD Last modified: 01-22-2002 Updater: julia

Definition: Heads of Europol National Units

POS: Noun Subject Area: Goverment & Public Authorities Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father unit; social unit

Term Table:

🇩🇰	HENU	Heads of Europol Natio...	lederne af de nationale ...	
🇳🇱	Hoofden van Europol N...			
🇬🇧	Heads of Europol Natio...	HENU	HENU group	
🇫🇮	Europolin kansallisten y...	Heads of Europol Natio...		
🇫🇷	chefs d'unités nationale...			
🇩🇪	Leiter der Nationalen E...			
🇬🇷	αρχηγοί εθνικών μονάδω...			
🇮🇹	Capi delle Unità nazion...	HENU		
🇵🇹	Chefes das Unidades ...			
🇪🇸	Cuerpo superior de Eur...	HENU		
🇸🇪	Ledarna för Europols n...	HENU		

Sense: 1 Entry Date: 10-10-2001 Originator: source Last modified: 12-07-2001 Updater: manuela

Gender: masculine Status: verified Source:

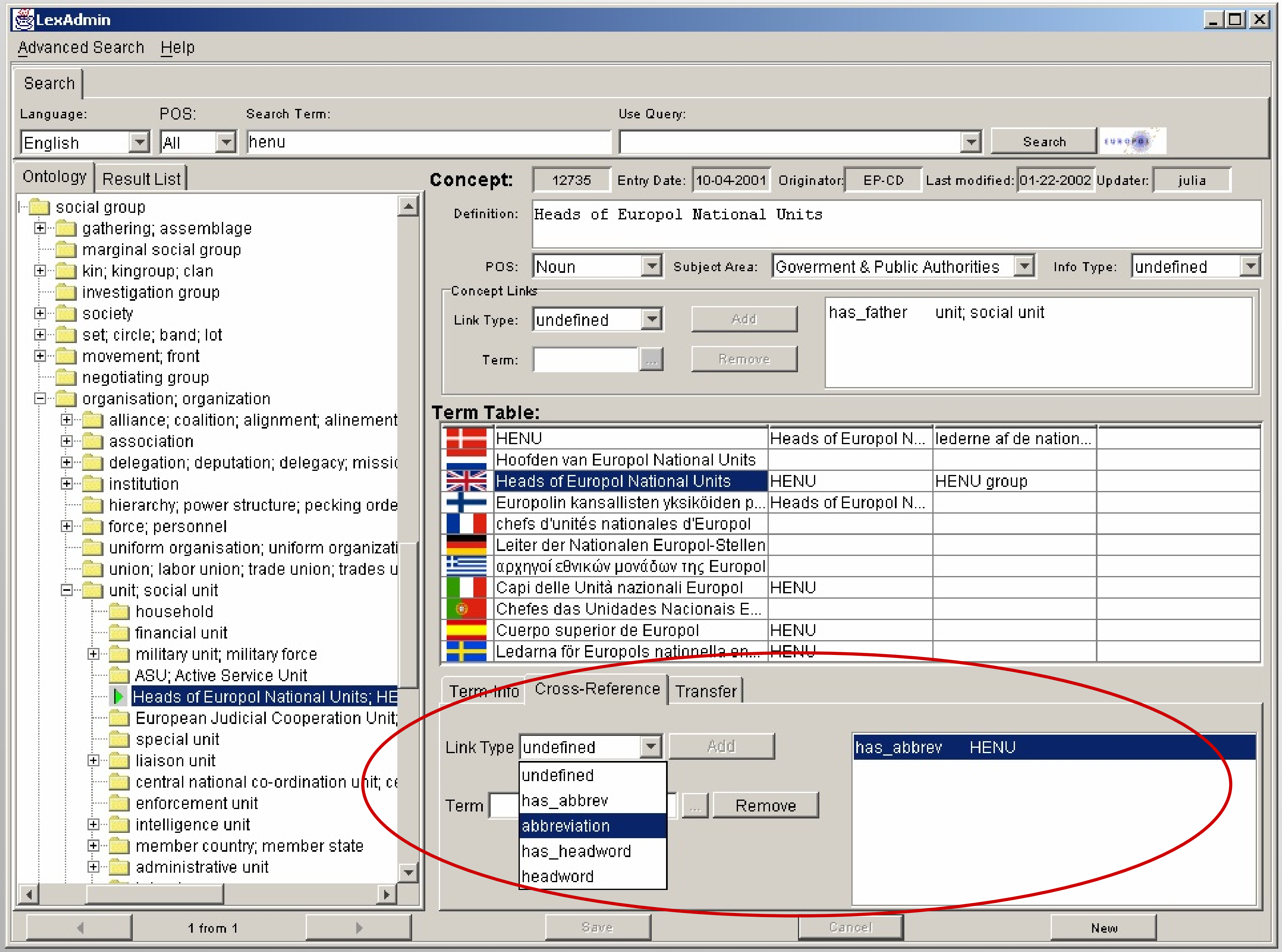
Number: Pluraletantum Entry Type: Multiword Usage: undefined

Transitivity: undefined MW Head: Capi Locale: undefined

Context:

- undefined
- Capi
- delle
- Unità
- nazionali
- Europol

Term Table: Cross Reference



Language: English POS: All Search Term: henu Use Query:

- social group
  - gathering; assemblage
  - marginal social group
  - kin; kingroup; clan
  - investigation group
  - society
  - set; circle; band; lot
  - movement; front
  - negotiating group
  - organisation; organization
    - alliance; coalition; alignment; alinement
    - association
    - delegation; deputation; delegacy; missic
    - institution
    - hierarchy; power structure; pecking orde
    - force; personnel
    - uniform organisation; uniform organizati
    - union; labor union; trade union; trades u
    - unit; social unit
      - household
      - financial unit
      - military unit; military force
      - ASU; Active Service Unit
      - Heads of Europol National Units; HE
      - European Judicial Cooperation Unit;
      - special unit
      - liaison unit
      - central national co-ordination unit; ce
      - enforcement unit
      - intelligence unit
      - member country; member state
      - administrative unit

Concept: 12735 Entry Date: 10-04-2001 Originator: EP-CD Last modified: 01-22-2002 Updater: julia

Definition: Heads of Europol National Units

POS: Noun Subject Area: Goverment & Public Authorities Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father unit; social unit

Term Table:

HENU	Heads of Europol N...	lederne af de nation...	
Hoofden van Europol National Units			
Heads of Europol National Units	HENU	HENU group	
Europolin kansallisten yksiköiden p...	Heads of Europol N...		
chefs d'unités nationales d'Europol			
Leiter der Nationalen Europol-Stellen			
αρχηγοί εθνικών μονάδων της Europol			
Capi delle Unità nazionali Europol	HENU		
Chefes das Unidades Nacionais E...			
Cuerpo superior de Europol	HENU		
Ledarna för Europols nationella en...	HENU		

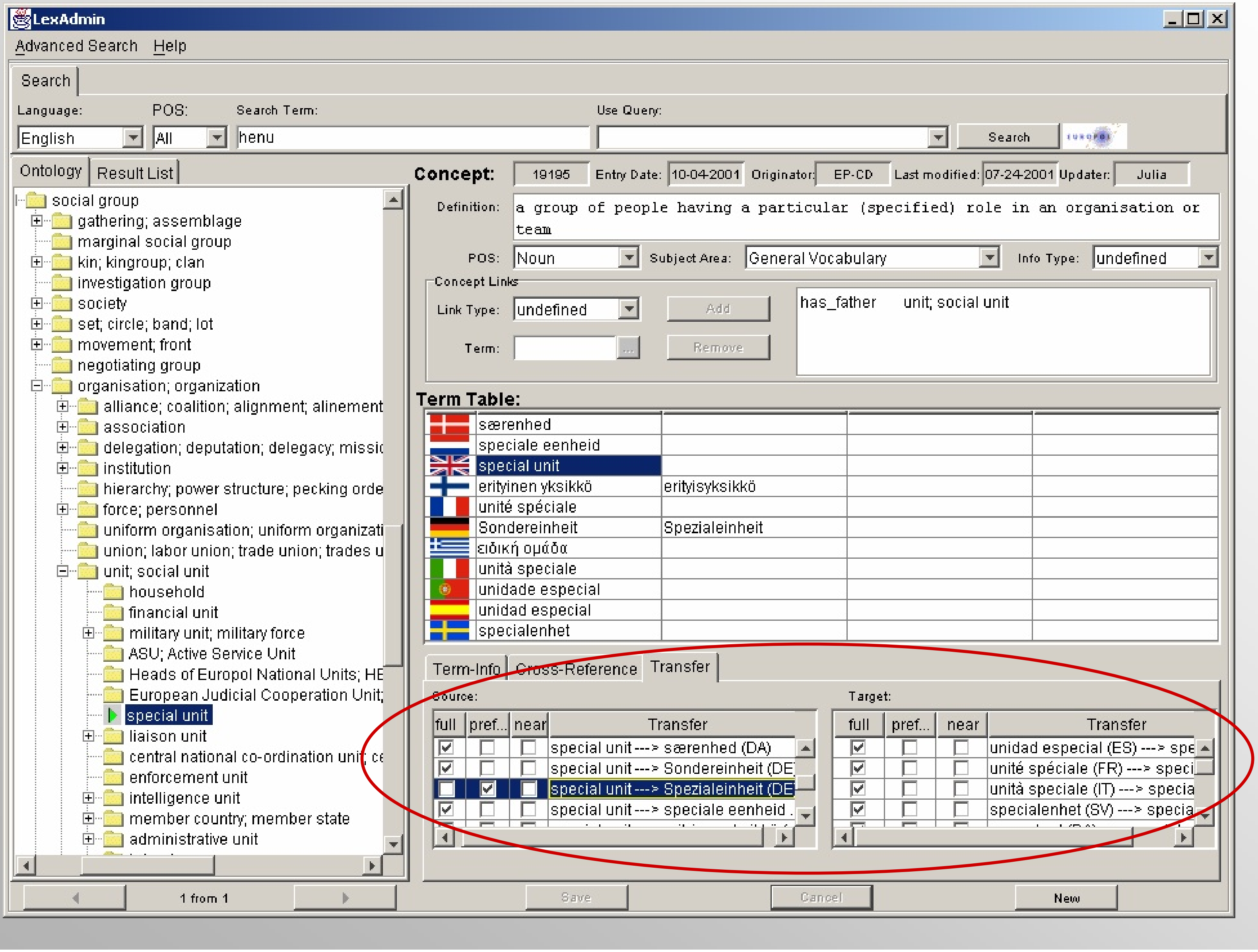
Link Type undefined

- undefined
- has_abbrev
- abbreviation
- has_headword
- headword

Term

has_abbrev HENU

Term Table: Transfers



Language: English POS: All Search Term: henu Use Query:

- social group
  - gathering; assemblage
  - marginal social group
  - kin; kingroup; clan
  - investigation group
  - society
  - set; circle; band; lot
  - movement; front
  - negotiating group
  - organisation; organization
    - alliance; coalition; alignment; alinement
    - association
    - delegation; deputation; delegacy; missic
    - institution
    - hierarchy; power structure; pecking orde
    - force; personnel
    - uniform organisation; uniform organizati
    - union; labor union; trade union; trades u
    - unit; social unit
      - household
      - financial unit
      - military unit; military force
      - ASU; Active Service Unit
      - Heads of Europol National Units; HE
      - European Judicial Cooperation Unit;
      - special unit
      - liaison unit
      - central national co-ordination unit; ce
      - enforcement unit
      - intelligence unit
      - member country; member state
      - administrative unit

Concept: 19195 Entry Date: 10-04-2001 Originator: EP-CD Last modified: 07-24-2001 Updater: Julia

Definition: a group of people having a particular (specified) role in an organisation or team

POS: Noun Subject Area: General Vocabulary Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father unit; social unit

Term Table:

særenhed			
speciale eenheid			
special unit			
erityinen yksikkö	erityisyksikkö		
unité spéciale			
Sondereinheit	Spezialeinheit		
ειδική ομάδα			
unità speciale			
unidade especial			
unidad especial			
specialenhet			

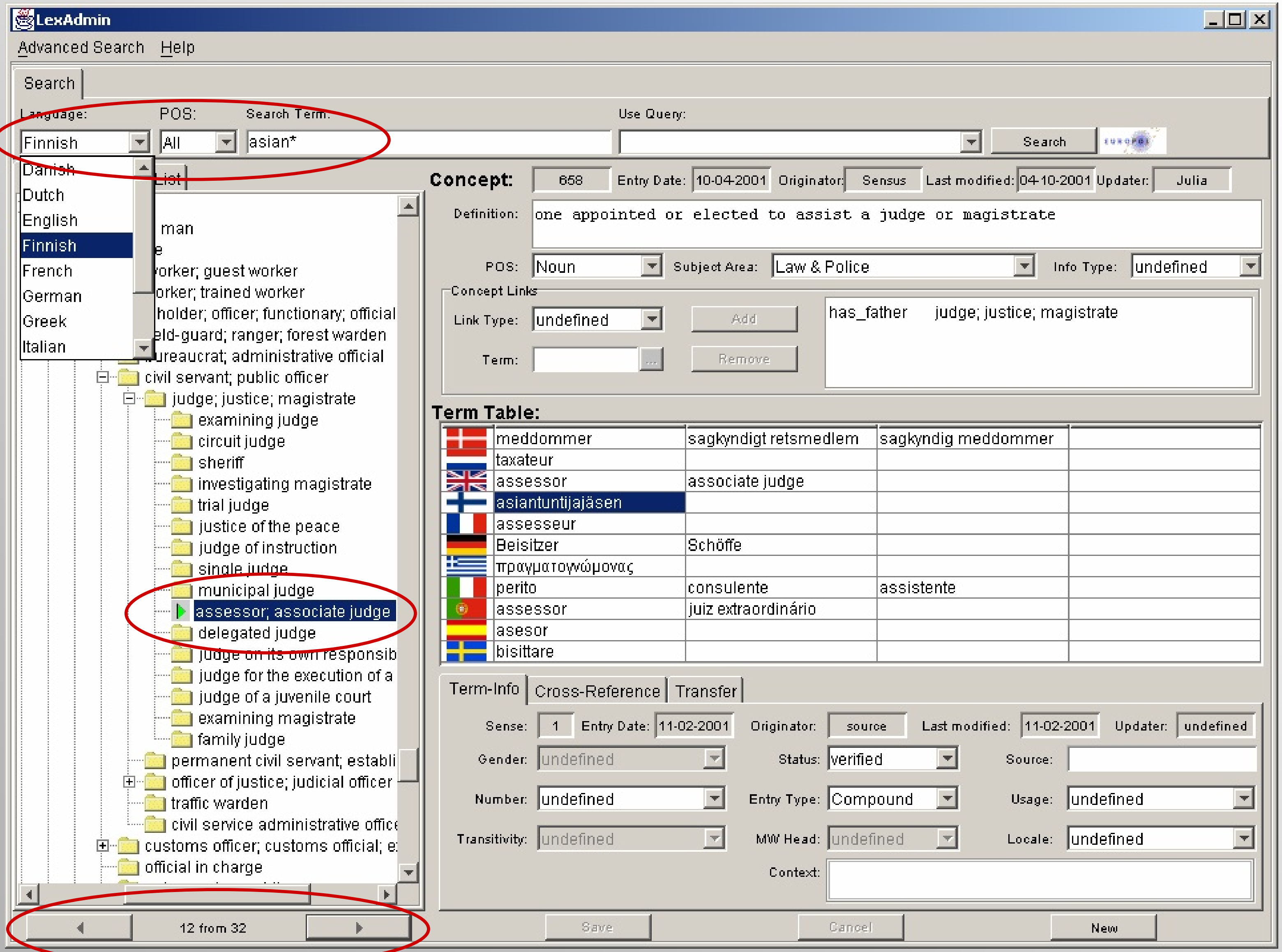
Source:

full	pref...	near	Transfer
☑	□	□	special unit ---> særenhed (DA)
☑	□	□	special unit ---> Sondereinheit (DE
□	☑	□	special unit ---> Spezialeinheit (DE
☑	□	□	special unit ---> speciale eenheid .

Target:

full	pref...	near	Transfer
☑	□	□	unidad especial (ES) ---> spe
☑	□	□	unité spéciale (FR) ---> speci
☑	□	□	unità speciale (IT) ---> specia
☑	□	□	specialenhet (SV) ---> specia

Search – Result in Ontology Tree



Language: Finnish
POS: All
Search Term: asian*
Use Query:

man
e
vorker; guest worker
orker; trained worker
holder; officer; functionary; official
eld-guard; ranger; forest warden
ureaucrat; administrative official
civil servant; public officer
judge; justice; magistrate
examining judge
circuit judge
sheriff
investigating magistrate
trial judge
justice of the peace
judge of instruction
single judge
municipal judge
assessor; associate judge
delegated judge
judge on its own responsib
judge for the execution of a
judge of a juvenile court
examining magistrate
family judge
permanent civil servant; establi
officer of justice; judicial officer
traffic warden
civil service administrative office
customs officer; customs official; e
official in charge

12 from 32

Concept: 658 Entry Date: 10-04-2001 Originator: Sensus Last modified: 04-10-2001 Updater: Julia

Definition: one appointed or elected to assist a judge or magistrate

POS: Noun Subject Area: Law & Police Info Type: undefined

Concept Links

Link Type: undefined
Term:

has_father judge; justice; magistrate

Term Table:

🇩🇰	meddommer	sagkyndigt retsmedlem	sagkyndig meddommer	
	taxateur			
🇬🇧	assessor	associate judge		
🇫🇮	asiantuntijajäsen			
🇫🇷	assesseur			
🇩🇪	Beisitzer	Schöffe		
🇬🇷	πραγματογνώμονας			
🇮🇹	perito	consulente	assistente	
🇵🇹	assessor	juiz extraordinário		
🇪🇸	asesor			
🇸🇪	bisittare			

Sense: 1 Entry Date: 11-02-2001 Originator: source Last modified: 11-02-2001 Updater: undefined

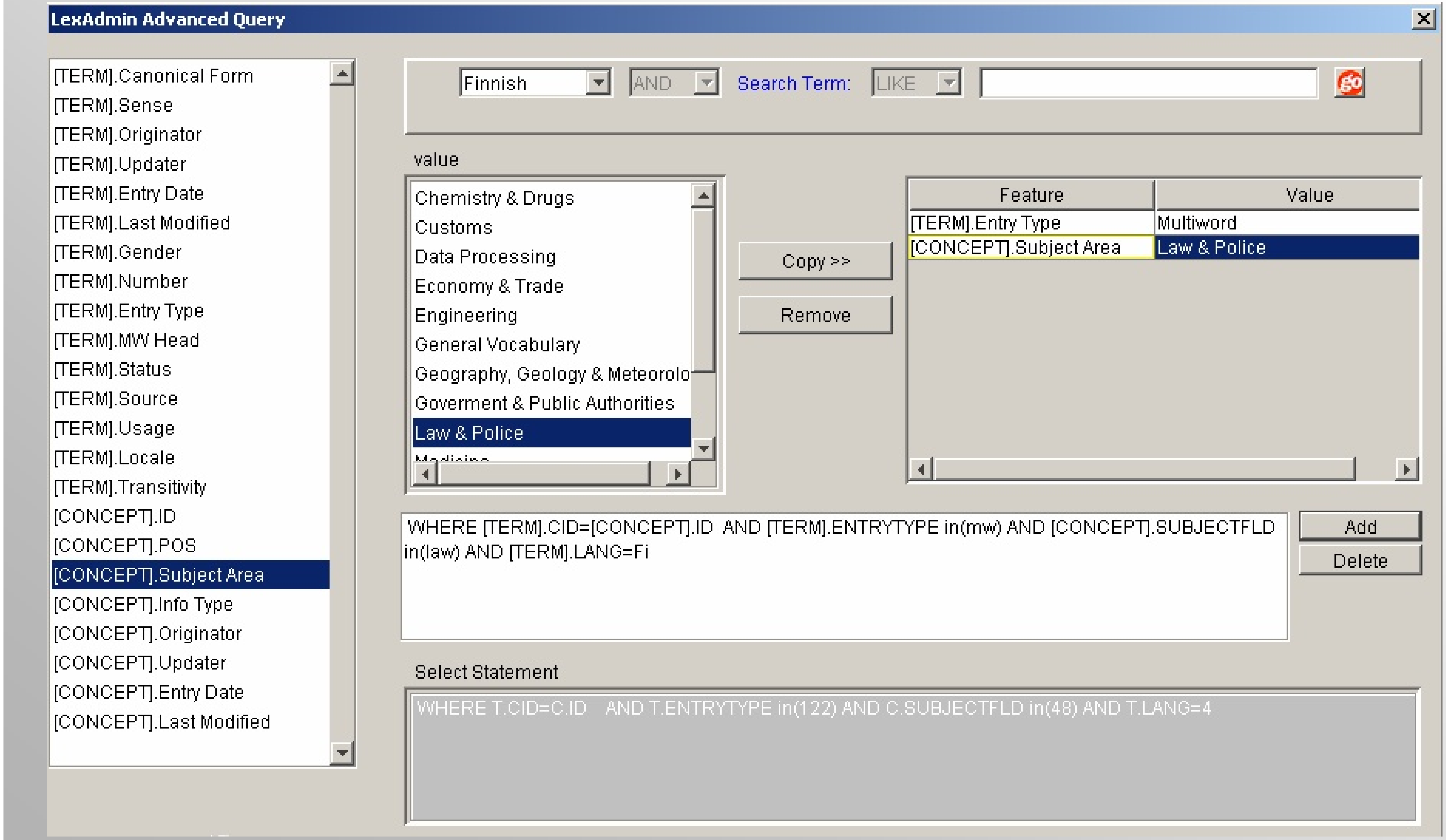
Gender: undefined Status: verified Source:

Number: undefined Entry Type: Compound Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Advanced Query Dialogue



Finnish | AND | Search Term: LIKE

[TERM].Canonical Form
[TERM].Sense
[TERM].Originator
[TERM].Updater
[TERM].Entry Date
[TERM].Last Modified
[TERM].Gender
[TERM].Number
[TERM].Entry Type
[TERM].MW Head
[TERM].Status
[TERM].Source
[TERM].Usage
[TERM].Locale
[TERM].Transitivity
[CONCEPT].ID
[CONCEPT].POS
[CONCEPT].Subject Area
[CONCEPT].Info Type
[CONCEPT].Originator
[CONCEPT].Updater
[CONCEPT].Entry Date
[CONCEPT].Last Modified

value

Chemistry & Drugs
Customs
Data Processing
Economy & Trade
Engineering
General Vocabulary
Geography, Geology & Meteorolo
Goverment & Public Authorities
Law & Police
Medicine

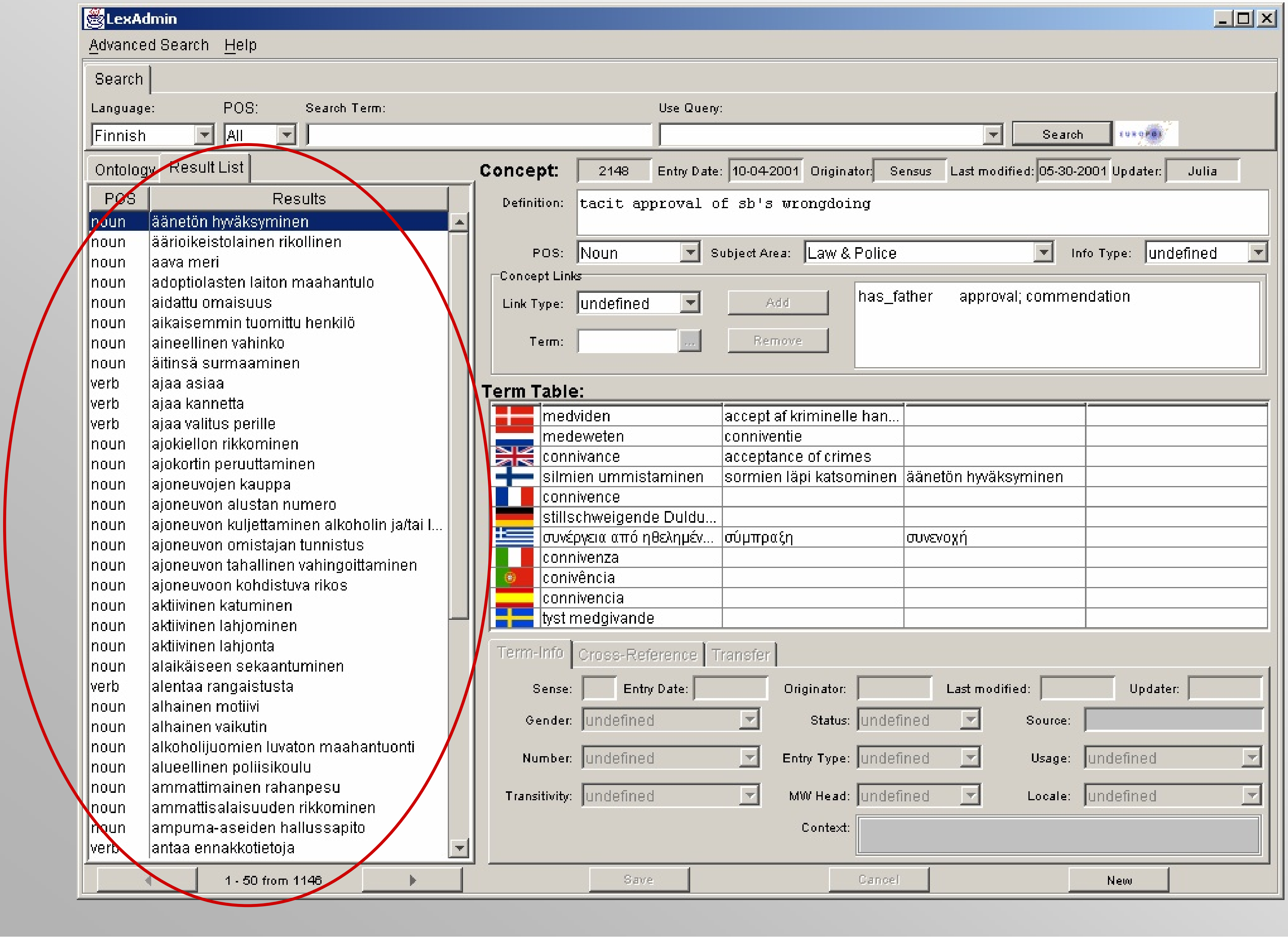
Feature	Value
[TERM].Entry Type	Multiword
[CONCEPT].Subject Area	Law & Police

```
WHERE [TERM].CID=[CONCEPT].ID  AND [TERM].ENTRYTYPE in(mw) AND [CONCEPT].SUBJECTFLD
in(law) AND [TERM].LANG=Fi
```

Select Statement

```
WHERE T.CID=C.ID    AND T.ENTRYTYPE in(122) AND C.SUBJECTFLD in(48) AND T.LANG=4
```

Advanced Query Result List



Language: Finnish
POS: All
Search Term:
Use Query:

POS	Results
noun	äänetön hyväksyminen
noun	äärioikeistolainen rikollinen
noun	aava meri
noun	adoptiolasten laiton maahantulo
noun	aidattu omaisuus
noun	aikaisemmin tuomittu henkilö
noun	aineellinen vahinko
noun	äitinsä surmaaminen
verb	ajaa asiaa
verb	ajaa kannetta
verb	ajaa valitus perille
noun	ajokiellon rikkominen
noun	ajokortin peruuttaminen
noun	ajoneuvojen kauppa
noun	ajoneuvon alustan numero
noun	ajoneuvon kuljettaminen alkoholin ja/tai l...
noun	ajoneuvon omistajan tunnistus
noun	ajoneuvon tahallinen vahingoittaminen
noun	ajoneuvoon kohdistuva rikos
noun	aktiivinen katuminen
noun	aktiivinen lahjominen
noun	aktiivinen lahjonta
noun	alaikäiseen sekaantuminen
verb	alentaa rangaistusta
noun	alhainen motiivi
noun	alhainen vaikutin
noun	alkoholijuomien luvaton maahantuonti
noun	alueellinen poliisikoulu
noun	ammattimainen rahanpesu
noun	ammattisalaisuuden rikkominen
noun	ampuma-aseiden hallussapito
verb	antaa ennakkotietoja

1 - 50 from 1146

Concept: 2148 Entry Date: 10-04-2001 Originator: Sensus Last modified: 05-30-2001 Updater: Julia

Definition: tacit approval of sb's wrongdoing

POS: Noun Subject Area: Law & Police Info Type: undefined

Concept Links

Link Type: undefined

Term:

has_father approval; commendation

Term Table:

medviden	accept af kriminelle han...	
medeweten	conniventie	
connivance	acceptance of crimes	
silmien ummistaminen	sormien läpi katsominen	äänetön hyväksyminen
connivence		
stillschweigende Duldu...		
συνέργεια από ηθελημέν...	σύμπραξη	συνενοχή
connivenza		
conivência		
connivencia		
tyst medgivande		

Sense: Entry Date: Originator: Last modified: Updater:

Gender: undefined Status: undefined Source:

Number: undefined Entry Type: undefined Usage: undefined

Transitivity: undefined MW Head: undefined Locale: undefined

Context:

Concept Manager: Functionality

- **Concept level coding**
  - Add / remove concepts
  - Link concepts to each other
- **Term level coding**
  - Add / remove terms
- Linguistic level
  - Code linguistic features
  - Code translation information
  - Code crossreferences
- **Import entries**
  - OLIF import format
  - Unlinked folder
- **Delete nodes / trees**
  - Save subtrees in folder
- **Compile entries**
  - Export for query translation
  - Export for MT
- **Query the DB**
  - Simple queries
  - Extended Boolean queries

Tool Support

- This resource needs to be **administered**
  - Single point of administration!

  => ConceptNet Administration Tool

- This resource must **feed the applications**
  - Resources for linguistic corpus analysis
  - Machine Translation dictionaries
    - Concept oriented vs. Reading oriented organisation
    - Bilingual directed vs. Multilingual

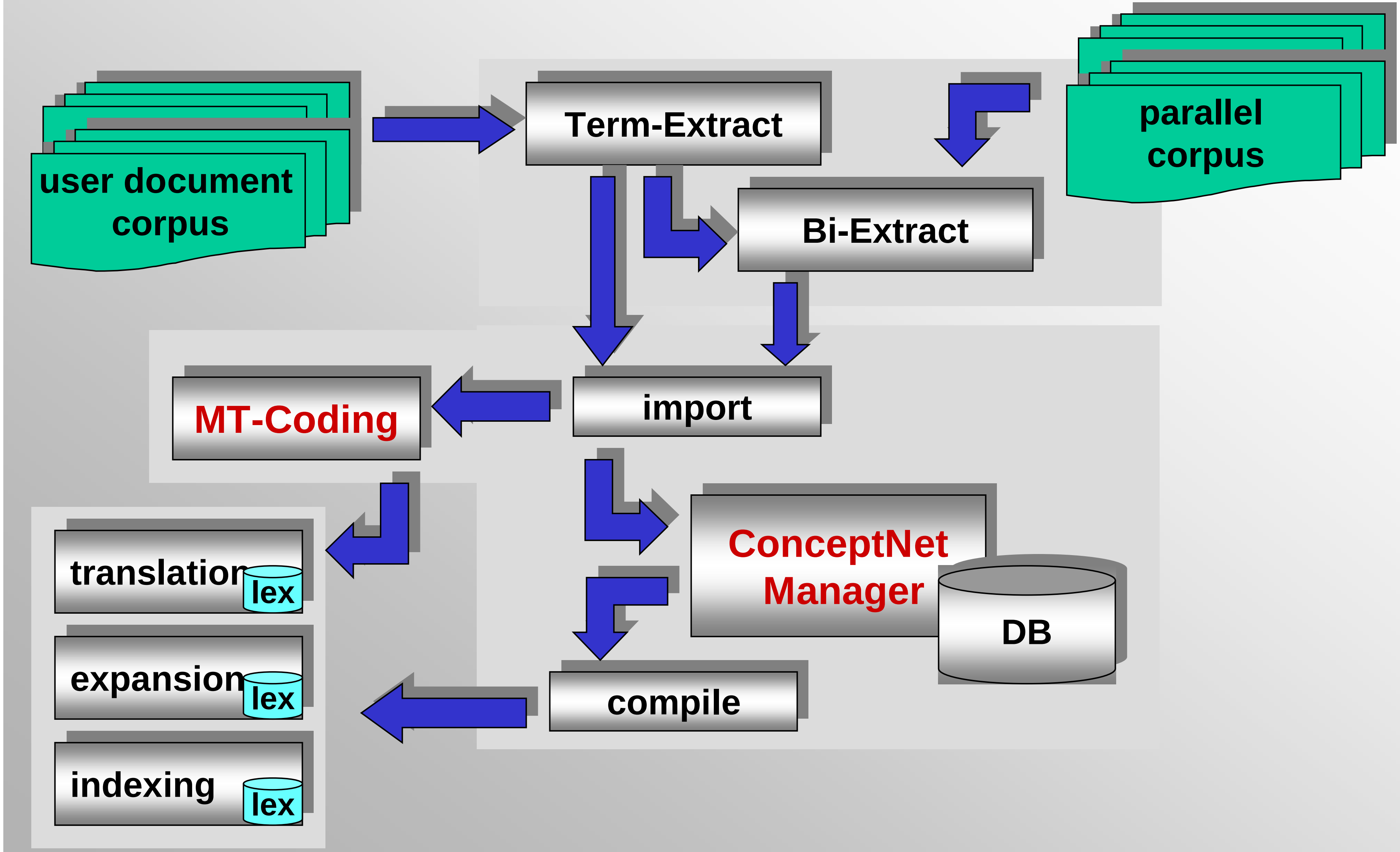
  => Exchange of (parts of) resources

    - Relevance of Exchange format (OLIF)

- This resource must be **set up**!
  - Because it is application-specific

  => Tools for corpus extraction and build-up

Resource Build-Up



More Intelligence ...

- Tools to create **linguistic annotations** for terms
  - Syntactic / semantic descriptions
- Tools to group terms into **concepts**
  - Monolingually (clustering, head-modifier-analysis)
- Tools to define **relations** between concepts
  - Hierarchical / non-hierarchical
- Tools to map **multilingual** concepts
  - (Extensions of current Bi-Extract)

(Business Constraint:

- Create a ConceptNet as described with 10K concepts in 3 languages in 2 weeks time)

Evaluation

- Only partial evaluations possible
- E.g.: IR using syntactic or semantic information:
  - POS Support
    - *Plays*$_{\text{-Noun}}$ *of Shakespeare* vs. *Child plays*$_{\text{-Verb}}$
    - No improvement due to performance of POS taggers in some language -> improve this first
  - Named Entity
    - Names like *Akrinski* in German texts are not mixed up
    - *Fisher*$_{\text{-person}}$ increases precision but reduces recall if NE is not safe enough

First improve what you already know is a source of deterioration

Thank you for your attention

g.thurmair@linguatec.de
www.linguatec.de