ConceptManager: A Tool to administer multilingual ~~Ontologies~~ ConceptNets

A. Jackson, M. Lewandowski
Gr. Thurmair, J. Zwickl

Outline

I The Task

II Solution Workflow

III ConceptManager Design

IV ConceptManager Features

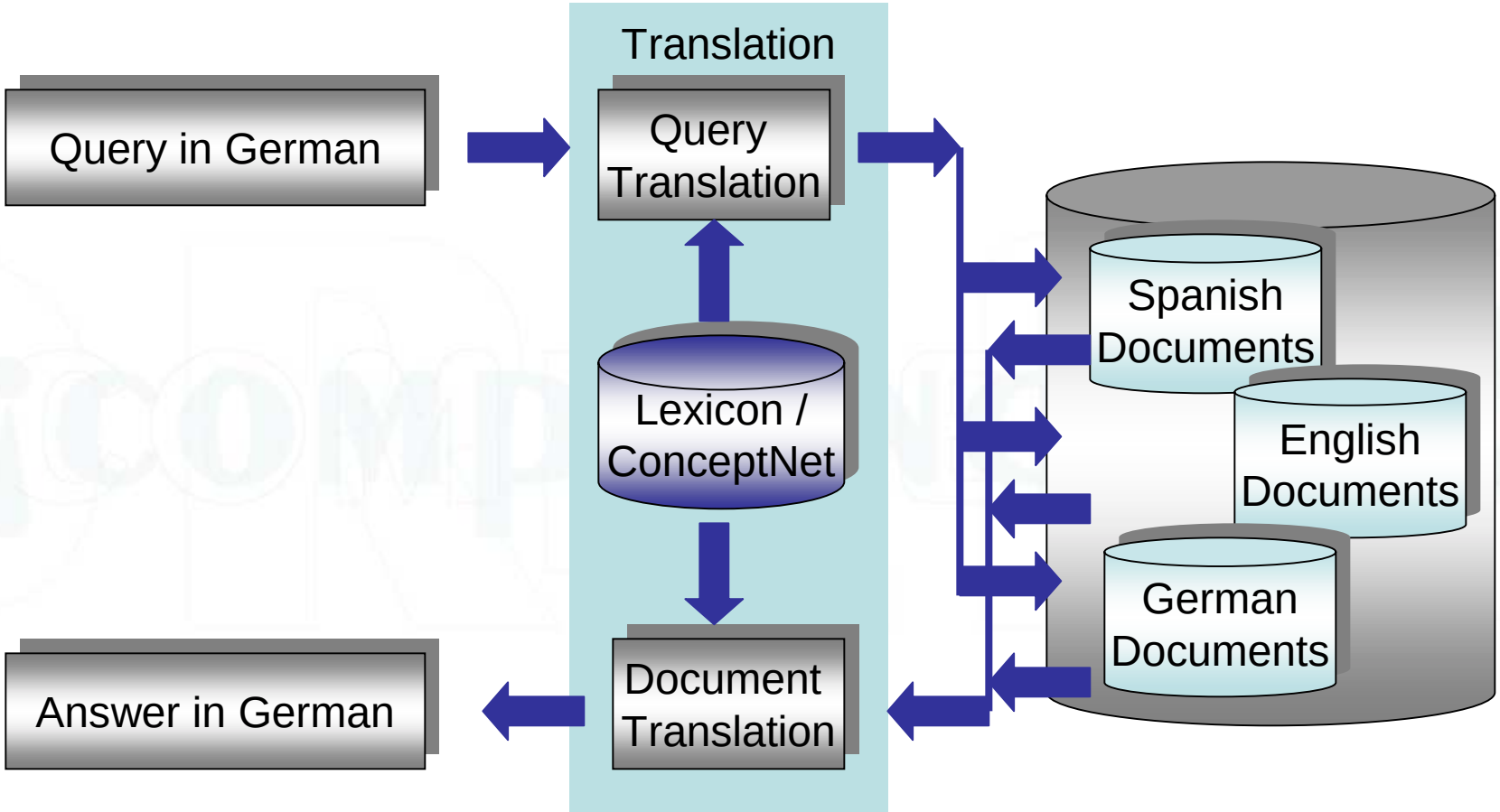
V Demo - Screens

I Task: Crosslingual Retrieval

Support of **Crosslingual Retrieval**

- **Indexing**:
  Provide linguistic knowledge for indexing
- **Query expansion**:
  Find alternative / better search terms
- **Query Translation**:
  Translate terms in target languages
- **Document re-translation**:
  Machine translation or key term translation

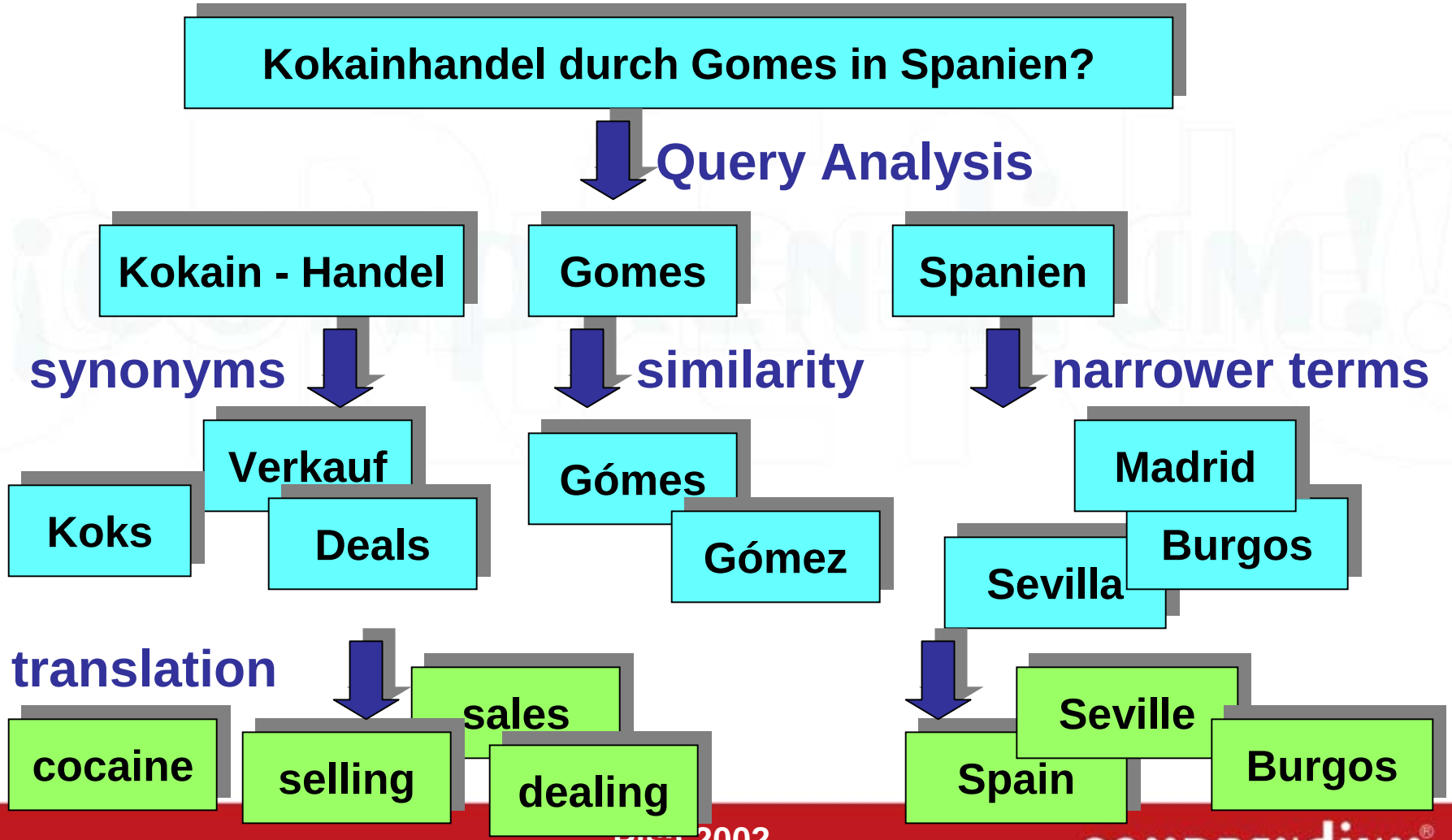
Task: Crosslingual Retrieval



Build linguistic resources for **query expansion** and **query and document translation**

Query Analysis, Expansion, Translation

Use **Multilingual ConceptNet**



Search result

dupin.doc

中国肃毒第一案

伍 岳

这是共和国成立以来最大的毒品走私案，也是我国破获的缴获毒品数量最多的一起贩毒案。国际警方将其列为世界第六、亚洲第一毒品案件。该宗案件代号“９６０１”，先后抓获涉案人员４２人，缴获含量为７６．２％的４号海洛因５９８．８５公斤，贩毒工具汽车１６辆，“中飞”快艇１艘，毒资６００多万元人民币和３１万元美元。

一、毒欲滔天

中国肃毒第一案如何发生、侦破、审理并判决呢？事情要从此案的关键人物，堪称“中国头号毒枭”的王世鉴说起。

王世鉴４０多岁，福建福鼎市人，虽说是“头号毒枭”，却并非人们想象中凶神恶煞、膀阔腰粗的模样。他中小个头，脸盘很窄，太阳穴和脑门部分肌肉干缩，面容瘦削中透出精明。他读书不多，初中毕业后即辍学随父做工，当过五金厂工人，后在老家福鼎做生意，１９９０年在广西凭祥经商。由于时运不济，王世鉴干什么亏什么，到１９９４年他累计欠债５００万元。

在如此窘迫的情况下，一筹莫展的王世鉴忽然想到贩毒可得暴利。１９９５年４月底，王世鉴携带１５０万元人民币到了缅甸的邦康，找到一个名叫阿强的人，向他买下了４０多只海洛因（每只７００克），装在吉普车的暗格里，然后运回广州以每包９万元人民币的价格卖给了吴浩铭。

Key Term Translation (direct translation)

dupin.doc

中国（China）肃毒第一案（Schlag gegen Rauschgifthandel）

伍 岳

这是共和国成立以来最大的毒（Droge）品走私案，也是我国破获的缴获毒（Droge）品数量最多的一起贩毒（Droge）案。国际警方（internationale Polizei）将其列为世界第六、亚洲第一毒（Droge）品案件。该宗案件代号“９６０１”，先后抓获涉案人员（Betroffene）４２人，缴获含量为７６．２％的４号海洛因（Heroin）５９８．８５公斤（Kilogramm），贩毒（Droge）工具汽车（Transportauto）１６辆，“中飞”快艇（Transportschnellboot）１艘，毒（Droge）资６００多万元人民币和３１万元美元（zehn Tausende Dollar）。

一、毒（Droge）欲滔天

中国（China）肃毒第一案（Schlag gegen Rauschgifthandel）如何发生（Geschichte）、侦破（Entdeckung）、审理（Verhandlung）并判决（Urteil）呢？事情要从此案的关键人物，堪称“中国（China）头号毒枭（groesster Rauschgifthaendler）”的王世鉴（Wang Si-Jian）说起。

王世鉴（Wang Si-Jian）４０多岁（Jahre alt），福建福鼎市（Stadt Fu-Di, Provinz Fu-Jian）人，虽说是“头号毒枭（groesster Rauschgifthaendler）”，却并非人们想象中凶神恶煞、膀阔腰粗的模样。他中小个头，脸盘很窄，太阳穴和脑门部分肌肉干缩，面容瘦削中透出精明。他读书不多，初中毕业后即辍学随父做工，当过五金厂工人，后在老家福鼎做生意，１９９０年在广西凭祥经商。由于时运不济，王世鉴（Wang Si-Jian）干什么亏什么，到１９９４年他累计欠债（Schulden）５００万元。

在如此窘迫的情况下，一筹莫展的王世鉴（Wang Si-Jian）忽然想到贩毒（Droge）可得暴利。１９９５年４月底，王世鉴（Wang Si-Jian）携带１５０万元人民币到了缅甸（Burma, Birma）的邦康，找到一个名叫阿强的人，向他买下了４０多只海洛因（Heroin）（每只７００克），装在吉普车的暗格里，然后运回广州以每包９万元人民币的价格卖给了吴浩铭。

II Solution Path

- **Query** Processing
  - Use a query expansion and translation component
- **Document** Processing
  - Use machine translation
    (Full MT or key term translation)

=>

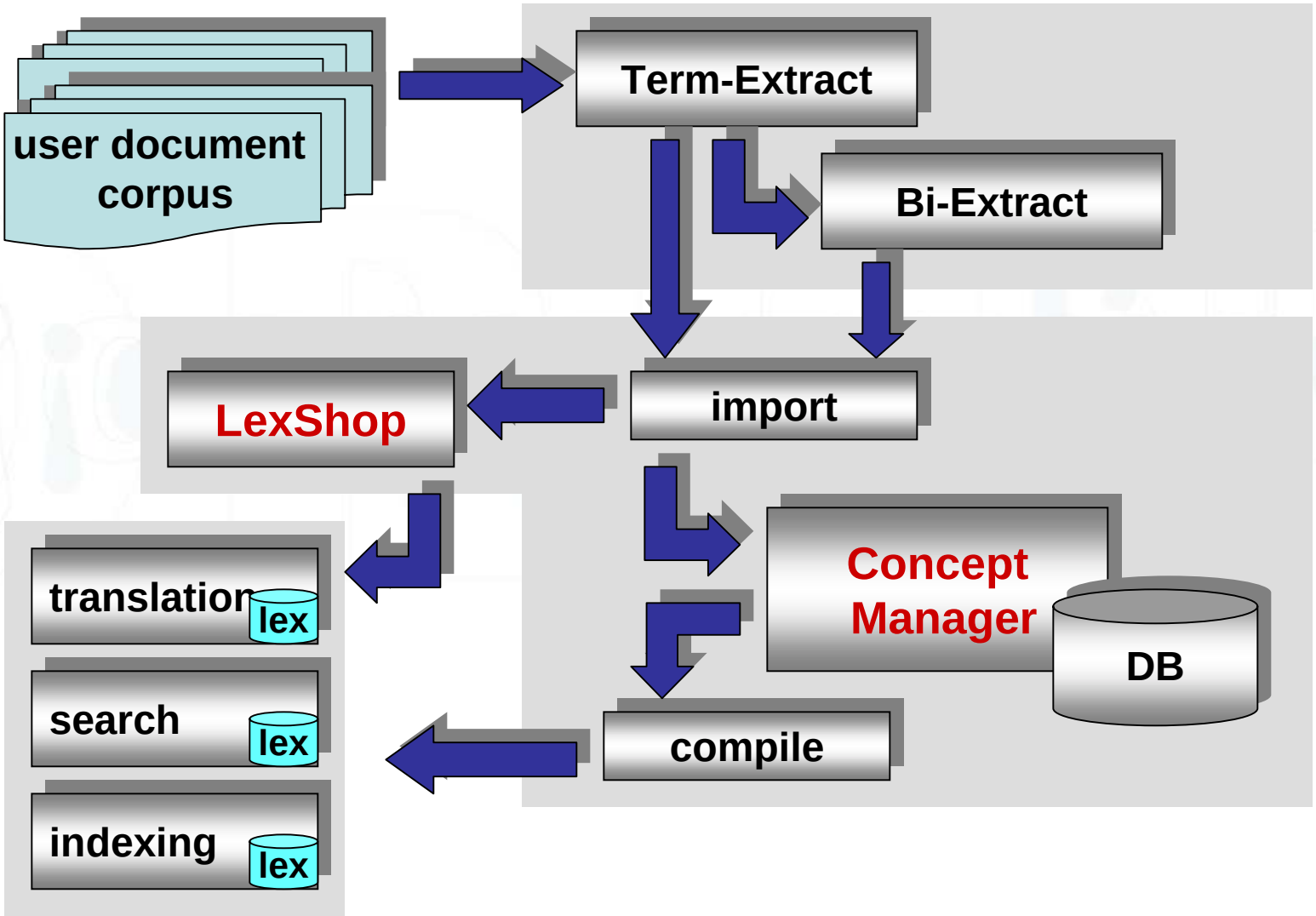
- Build linguistic resources for CLIR
- Start with users‘ existing material
  => Corpus analysis / corpus extraction

Building customised resources

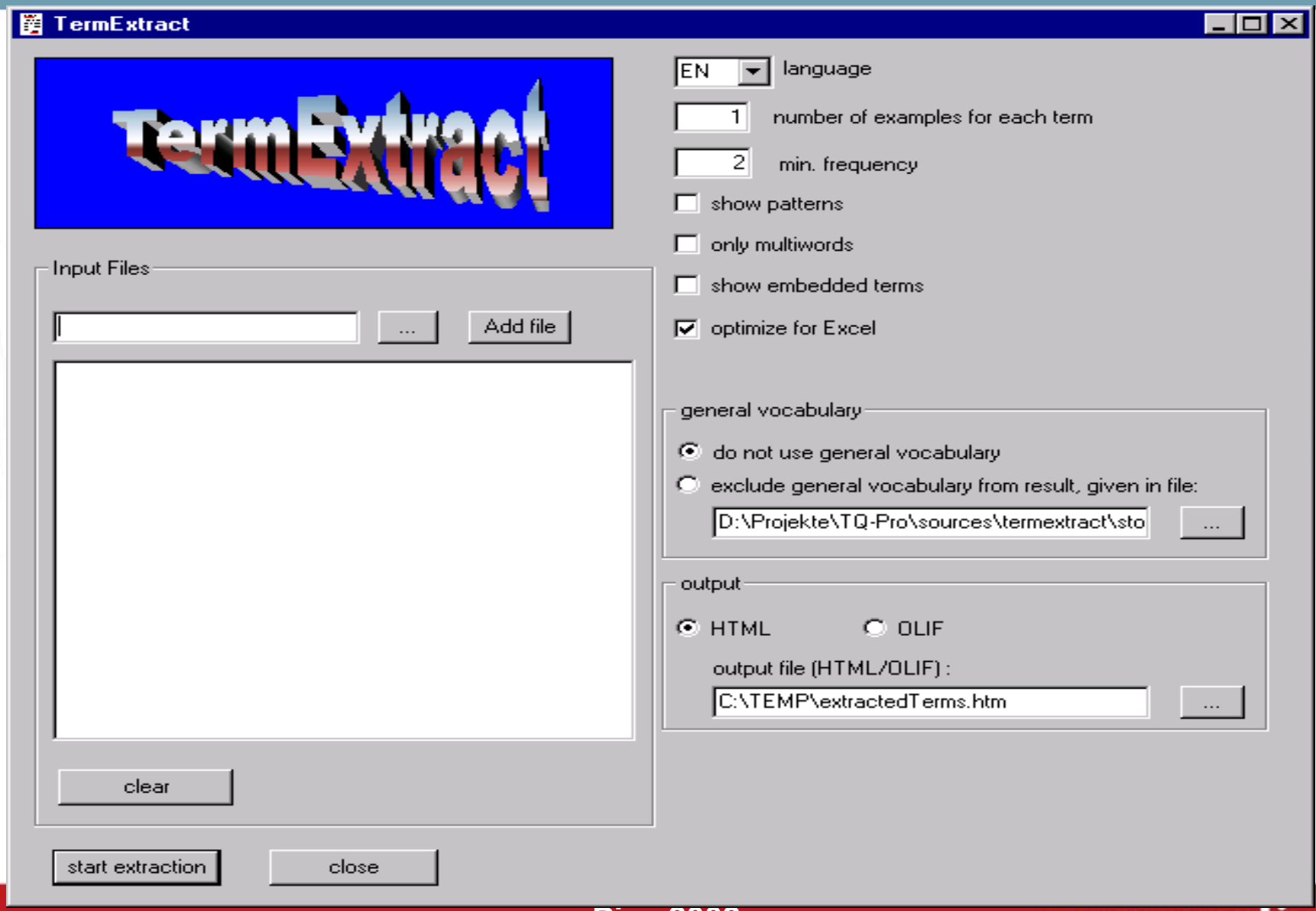
Building multilingual resources, using OLIF interface:

- Start from users‘ own material (corpus)
- Extract relevant terminology
  - Monolingual term extraction
  - bilingual extraction (based on ILSP-aligned parallel corpora)
- Machine Translation Import
  - **LexShop** as administration tool for MT lexicon
- Query and key term translation support
  - Import into multilingual ConceptNet
  - **ConceptManager** as administration tool
  - Compile into participating components
- But: **Single point of administration**

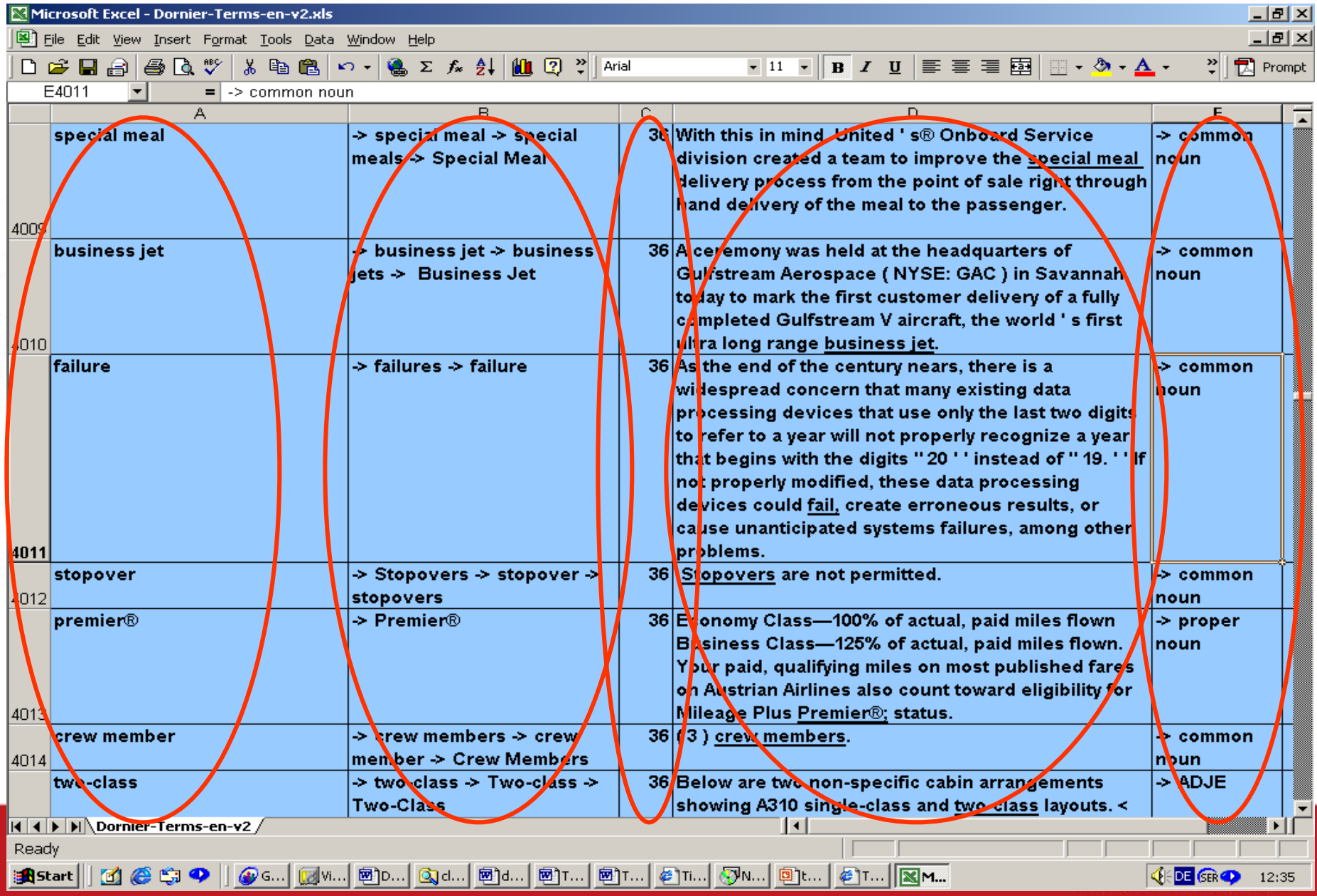
CLIR: Workflow



TermExtract



Beispiel: TermExtrakt



Microsoft Excel - Dornier-Terms-en-v2.xls

File Edit View Insert Format Tools Data Window Help

E4011 = -> common noun

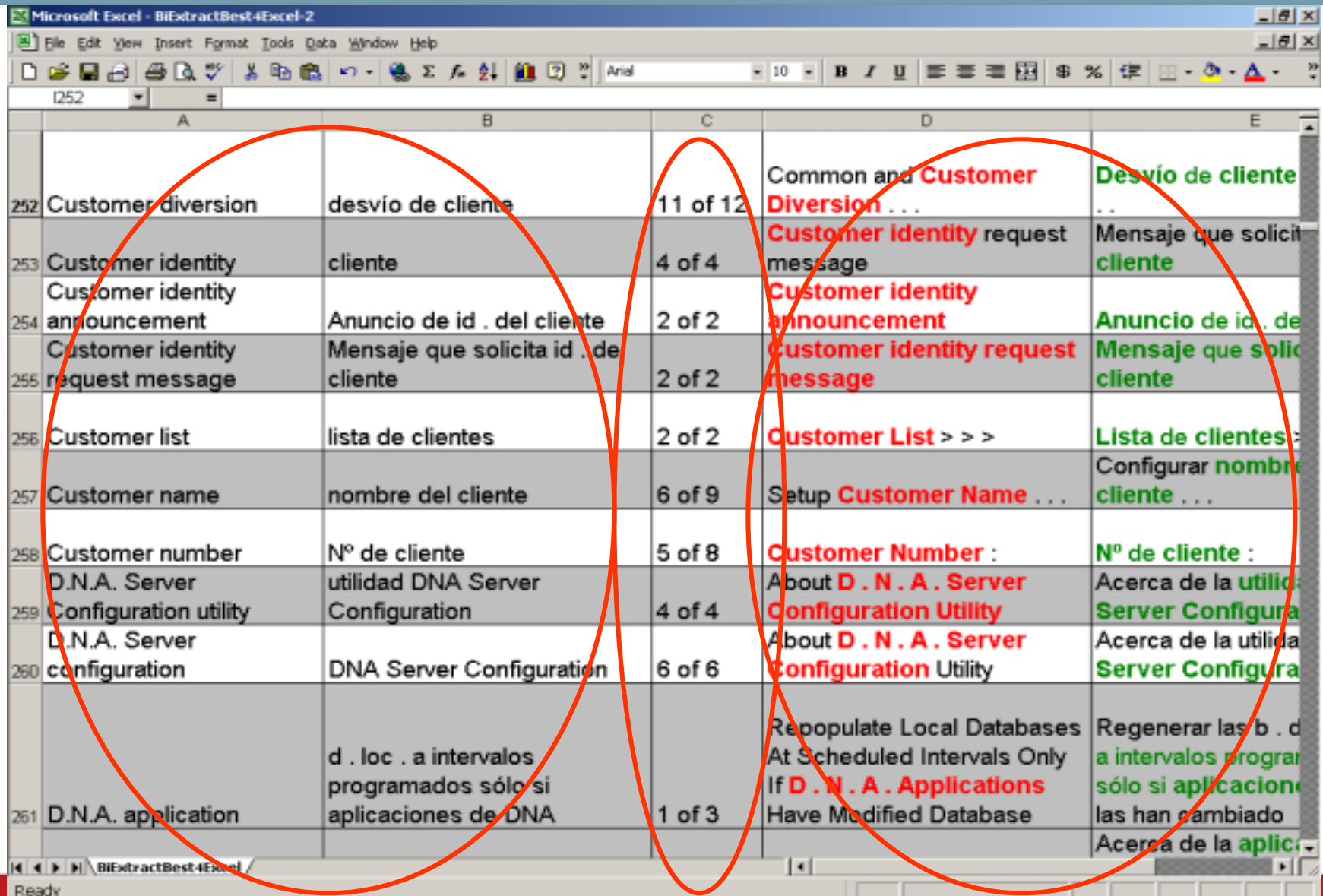
	A	B	C	D	E
4009	special meal	-> special meal -> special meals -> Special Meal	36	With this in mind, United ' s® Onboard Service division created a team to improve the special meal delivery process from the point of sale right through hand delivery of the meal to the passenger.	-> common noun
4010	business jet	-> business jet -> business jets -> Business Jet	36	A ceremony was held at the headquarters of Gulfstream Aerospace (NYSE: GAC) in Savannah today to mark the first customer delivery of a fully completed Gulfstream V aircraft, the world ' s first ultra long range business jet.	-> common noun
4011	failure	-> failures -> failure	36	As the end of the century nears, there is a widespread concern that many existing data processing devices that use only the last two digits to refer to a year will not properly recognize a year that begins with the digits " 20 ' ' instead of " 19. ' ' If not properly modified, these data processing devices could fail, create erroneous results, or cause unanticipated systems failures, among other problems.	-> common noun
4012	stopover	-> Stopovers -> stopover -> stopovers	36	Stopovers are not permitted.	-> common noun
4013	premier®	-> Premier®	36	Economy Class—100% of actual, paid miles flown Business Class—125% of actual, paid miles flown. Your paid, qualifying miles on most published fares on Austrian Airlines also count toward eligibility for Mileage Plus Premier®; status.	-> proper noun
4014	crew member	-> crew members -> crew member -> Crew Members	36	(3) crew members.	-> common noun
	two-class	-> two-class -> Two-class -> Two-Class	36	Below are two non-specific cabin arrangements showing A310 single-class and two-class layouts. <	-> ADJE

Dornier-Terms-en-v2

Ready

Start 12:35

BiExtract Output



Microsoft Excel - BiExtractBest4Excel-2

File Edit View Insert Format Tools Data Window Help

I252 =

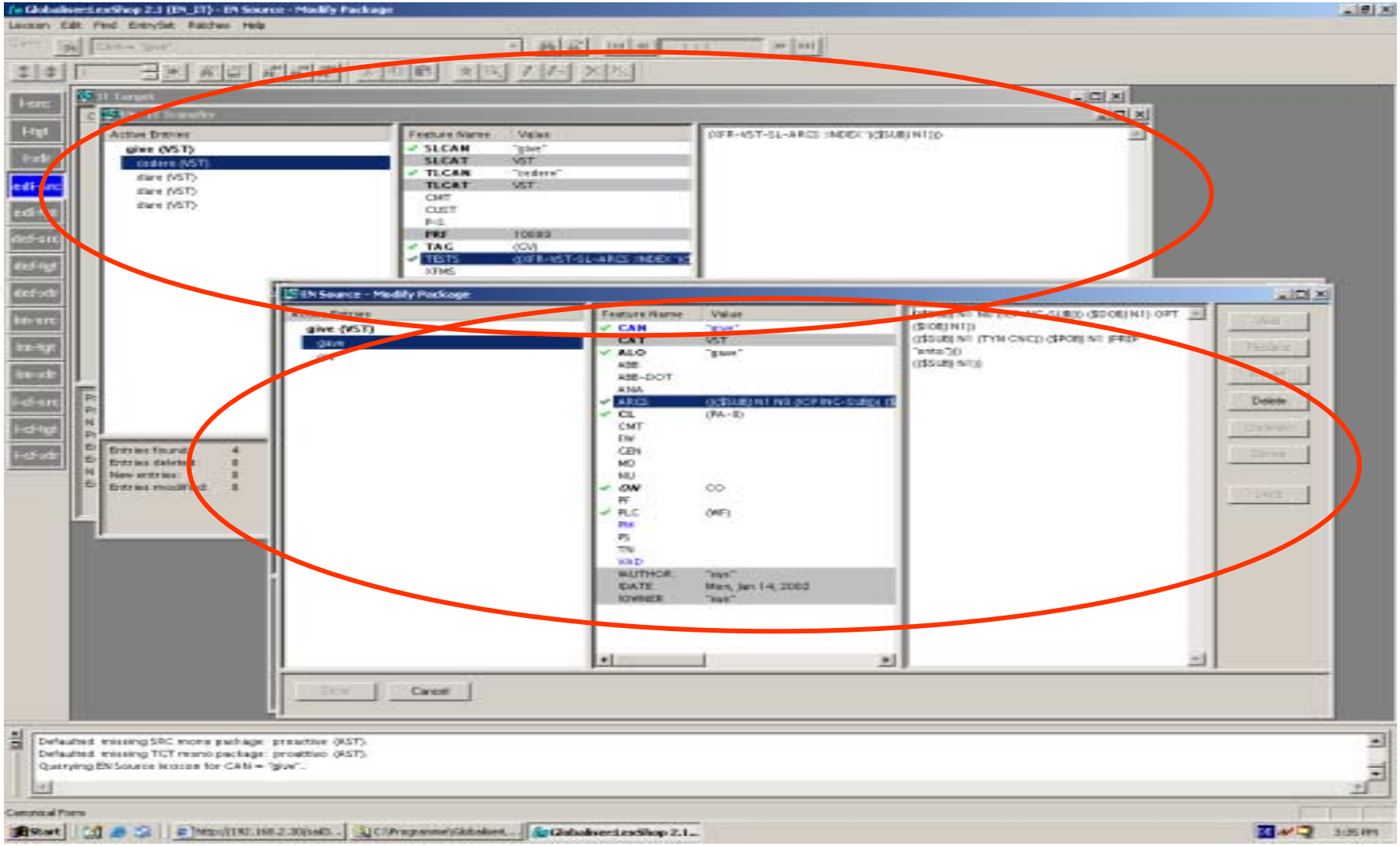
	A	B	C	D	E
252	Customer diversion	desvío de cliente	11 of 12	Common and Customer Diversion . . .	Desvío de cliente . .
253	Customer identity	cliente	4 of 4	Customer identity request message	Mensaje que solicit cliente
254	Customer identity announcement	Anuncio de id . del cliente	2 of 2	Customer identity announcement	Anuncio de id . de
255	Customer identity request message	Mensaje que solicita id . de cliente	2 of 2	Customer identity request message	Mensaje que solic cliente
256	Customer list	lista de clientes	2 of 2	Customer List > > >	Lista de clientes >
257	Customer name	nombre del cliente	6 of 9	Setup Customer Name . . .	Configurar nombr cliente . . .
258	Customer number	Nº de cliente	5 of 8	Customer Number :	Nº de cliente :
259	D.N.A. Server Configuration utility	utilidad DNA Server Configuration	4 of 4	About D . N . A . Server Configuration Utility	Acerca de la utilid Server Configura
260	D.N.A. Server configuration	DNA Server Configuration	6 of 6	About D . N . A . Server Configuration Utility	Acerca de la utilida Server Configura
261	D.N.A. application	d . loc . a intervalos programados sólo si aplicaciones de DNA	1 of 3	Repopulate Local Databases At Scheduled Intervals Only If D . N . A . Applications Have Modified Database	Regenerar las b . d a intervalos prograr sólo si aplicacion las han cambiado
					Acerca de la aplic

BiExtractBest4Excel

Ready

Start | Term... | Greg... | Visu... | Mark... | Meet... | Meet... | ~72... | VW-... | TER... | WinZ... | BiEx... | BiExt... | 12:15 PM

LexShop: MT Coding Tool



III Requirements for ConceptManager

- Must support query **expansion**
  - Build hierarchy / ontology of **concepts**
  - Cover most frequently used **terms**
- Must support term **translation**
  - For query and key term document translation
  - Translations on concept level (precision)
- Must support **linguistic components**
  - Compilers for machine translation, info extraction
  - Provide required linguistic attributes

Such a resource and tool did not exist

- **Lexicon** tools and term banks do not support ontologies
- **Ontology** editors do not support linguistic attributes

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

	WordNet	ConceptNet
Part-of-speech	obligatory	obligatory
Gloss	optional	obligatory
Example	optional	optional **term level**
Domain	(no)	obligatory

- Multilingual
  - one generic hierarchy (with no terms for some languages)
- Linguistic Annotations per term
- Relations between synset terms
  - On transfer level: preferred / (tests&actions)
  - On crossreference level: abbreviations, headwords, …
- WordNet is not a WordNet

Are ConceptNets domain-independent?

- Multilingual ConceptNet is **domain specific**
  - 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

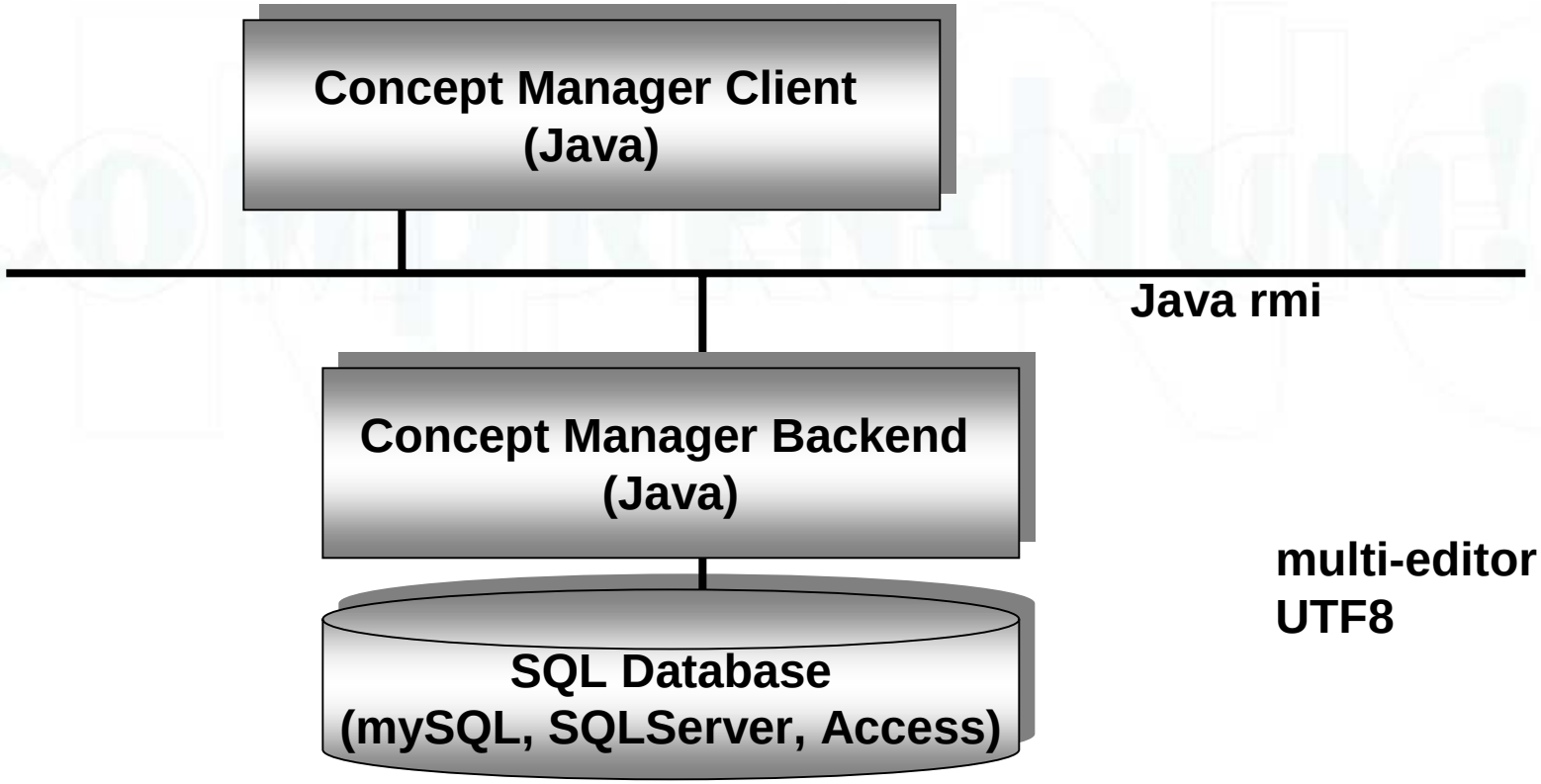
Some concepts do not have terms for some languages

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**
  - Cover all relevant terms in the domain
- Enrich concepts with **multilingual** terms
  - 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

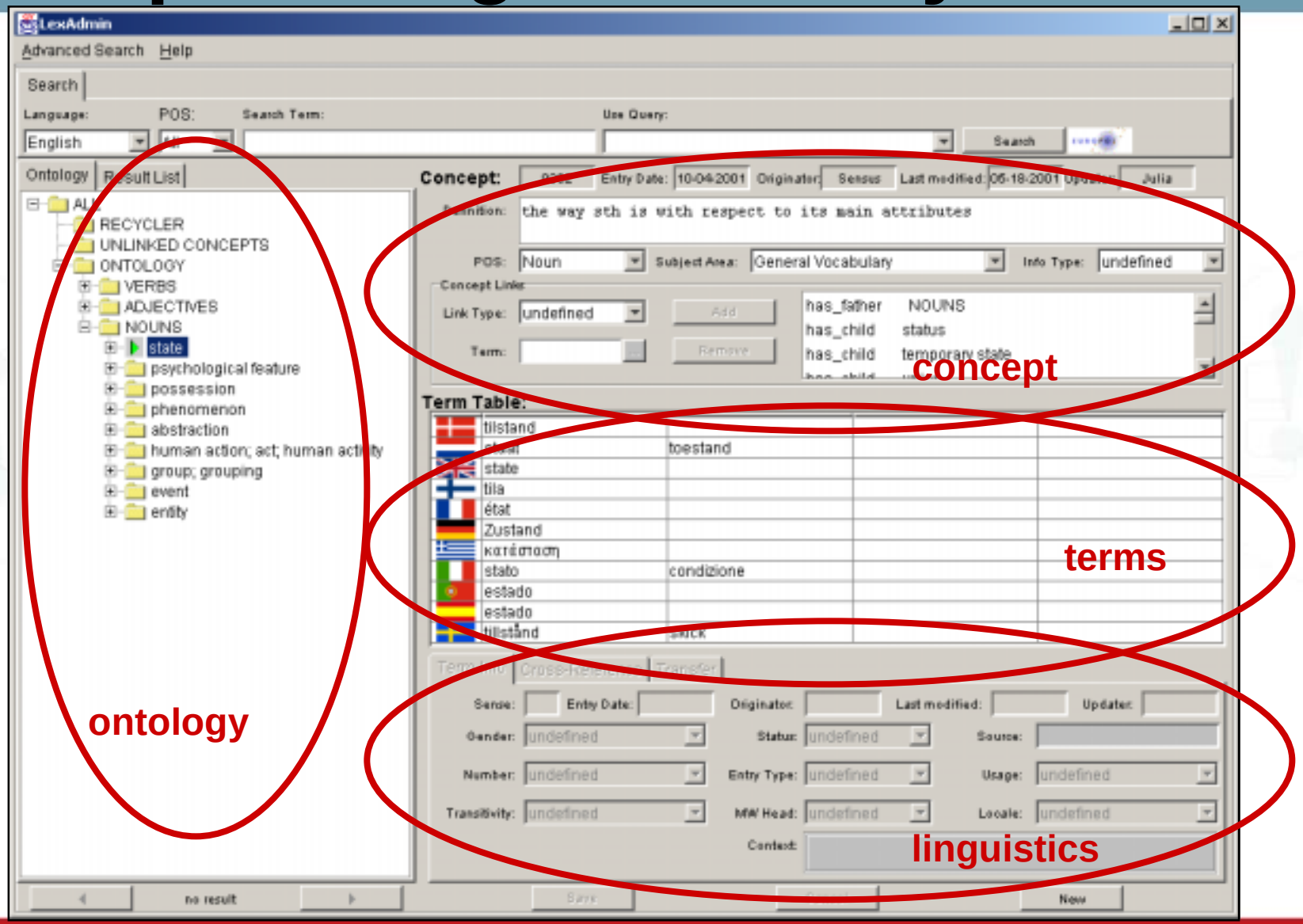
IV ConceptManager: Software Structure

Software Structure



multi-editor
UTF8

Concept Manager: GUI Layout



III Concept Manager: Functionality

- Concept level coding
  - Add / remove concepts
  - Link concepts to each other
- Term level coding
  - Add / remove terms
- Linguistic level coding
  - 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

Building a customer ConceptNet

- Building the hierarchy (English as pivot)
  - TermExtract their material (10K concepts)
  - Import into ConceptManager (unlinked concepts)
  - Create ontology nodes and links (16K concepts)
    - Nodes for systematic reasons
    - Nodes which are not terms
- Adding the other 10 languages
  - Align their material into memories (ILSP)
  - BiExtract their material
  - Use as proposals for translators
- Quality assurance and cross-checks
  - Criterion: the CLIR application
- Result: 16K concepts, ~ 230 K terms in 11 languages