Computational Dictionaries & Terminology

February 1 and 6, 2006

Dr. Andreas Eisele
Computerlinguistik & DFKI
eisele@dfki.de

Language Technology I
WS 2005/2006

Computational Dictionaries & Terminology

UNIVERSITÄT DES SAARLANDES

- Motivation
- Definitions
- Relevant Standards
- Important Resources
- Automatic Acquisition
- Outlook

Motivation for computational dictionaries

UNIVERSITÄT DES SAARLANDES

Natural language processing needs knowledge about words

- Morphological behavior (how words look/are modified)
- Syntactic behavior (part-of-speech, relation to other words)
- Semantic behavior (how words relate to meanings)

But: Construction of lexical resources is a major investment

- Vocabularies can be very large, e.g. Duden's ***Deutsche Rechtschreibung*** ~ 115000, the ***Oxford English Dictionary*** (***OED***) ~ 301100, ***Der Große Muret Sanders*** ~ 560000 entries, technical terminologies may contain millions of entries
- Words can have many meanings, can be used in multiple ways, e.g. entry for "get" in ***Der Kleine Muret Sanders:*** ~ 340 lines
- Selection of vocabulary may depend on application, subsets may be almost unlimited in size (person/place/company names), and may change quickly over time (product names, computer jargon)
- Theory-specific description of syntactic/semantic behavior makes re-use difficult

Motivation for terminology

UNIVERSITÄT DES SAARLANDES

…terminology as a structured set of concepts and their designations in a particular subject field… can be considered the infrastructure of specialized knowledge. Technical writing and technical documentation are impossible without properly using terminological resources. High-quality multilingual terminologies have become scarce and much desired commodities for language and knowledge industries. [Galinski/Budin '96]

Research field: terminology science

"the scientific study of the concepts and terms found in special languages" [ISO 1087]

Practical field of application: terminology management

- creation of subject-field specific terminologies
- recording in terminology databases, dictionaries, lexicons, specialized encyclopedias

Some definitions (from Wikipedia)

UNIVERSITÄT DES SAARLANDES

- A **dictionary** is a list of words with their definitions, a list of characters with their glyphs, or a list of words with corresponding words in other languages. In some languages, words can appear in many different forms, but only the lemma form appears as the main word or headword in most dictionaries. Many dictionaries also provide pronunciation information; grammatical information; word derivations, histories, or etymologies; illustrations; usage guidance; and examples in phrases or sentences.
- The word **thesaurus** … means a listing of words with similar, related, or opposite meanings (this meaning of *thesaurus* dates back to Roget's Thesaurus). For example, a book of jargon for a specialized field; or more technically a list of subject headings and cross-references used in the filing and retrieval of documents (or indeed papers, certificates, letters, cards, records, texts, files, articles, essays and perhaps even manuscripts), film, sound recordings, machine-readable media, etc. The first example of this genre, Roget's Thesaurus, was published in 1852, having been compiled earlier, in 1805, by Peter Roget. Entries in *Roget's Thesaurus* are not listed alphabetically but conceptually and are a great resource for writers.
- A **glossary** is a list of terms with the definitions for those terms. Traditionally, a glossary appears at the end a book and includes terms within that book which are either newly introduced or at least uncommon. In a more general sense, a glossary contains explanations of concepts relevant to a certain field of study or action. In this sense, the term is contemporaneously related to ontology.

More definitions

UNIVERSITÄT DES SAARLANDES

- **Terminology**, in its general sense, simply refers to the usage and study of terms — words and compound words generally used in specific contexts. The term "terminology" may also refer to a more formal discipline which systematically studies of the labelling or designating of concepts particular to one or more subject fields or domains of human activity, through research and analysis of terms in context, for the purpose of documenting and promoting correct usage. This study can be limited to one language or can cover more than one language at the same time (*multilingual terminology*, *bilingual terminology*, and so forth).
- In Information Science, an **ontology** is the product of an attempt to formulate an exhaustive and rigorous conceptual schema about a domain. An ontology is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (e.g., a **domain ontology**). However, computational ontology does not have to be hierarchical at all. The computer science usage of the term *ontology* is derived from the much older usage of the term ontology in philosophy.
- **WordNet** is a semantic lexicon for the English language. It groups English words into sets of synonyms called *synsets*, provides short definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The database and software tools … can be downloaded and used freely.

Some Important Resources

UNIVERSITÄT DES SAARLANDES

- **Celex** DB for Dutch, English, German: http://www.ru.nl/celex/
- **EAGLES** guidelines including computational lexicons: http://www.ilc.cnr.it/EAGLES96/browse.html
- **ELRA** catalogue (http://www.elda.org/rubrique2.html):
  - 57 monolingual lexicons
  - 41 bi- and multilingual lexicons
  - 22 terminological resources
- **Wordnet, EuroWordnet**
- **FrameNet, Propbank**
- **Eurodicautom (1973)…**

  … is the European Commission's multilingual term bank, based on phrasal automatic dictionary Dicautom (1964), and translation dictionary Euroterm (1962-68). Original languages: Dutch, French, German and Italian,added later: Danish,English (1973), Greek (1981), Portuguese and Spanish (1986), Finnish and Swedish (1995). Latin is also available.
- **IATE (in the process of replacing Eurodicautom since 2000)**

  … is the EU inter-institutional terminology database system. It will be used for the collection, dissemination and shared management of EU-specific terminology. This system will be multilingual and will be available to EU agencies and institutions, freelance translators and European citizens.
- **Eurovoc**

  …is a multilingual thesaurus covering the fields in which the European Communities are active; it provides a means of indexing the documents in the documentation systems of the European institutions and of their users. **Eurovoc 4.2** exists in 16 official languages but the missing ones will be added
- **IPC** (International Patent Classification): see separate slides

Wordnet – a semantic lexicon

UNIVERSITÄT DES SAARLANDES

Wordnet …

- …groups together words according to lexical semantic relations like *synonymy*, *hyponymy*, *meronymy*, *antonymy*, etc.
- …covers 4 open part-of-speech classes: nouns, verbs, adjectives, adverbs
- Words are assigned to sets of synonyms (synsets), all other relations hold between synsets
- WN has been used in many experiments on semantic disambiguation, IR, …
- WN has been ported to many other languages, attempts to build cross-lingual versions are on the way

Some Terminology Standards

UNIVERSITÄT DES SAARLANDES

Unfortunately, these standards refer to the data formats, not to the contents of terminology files

- **TBX:** Termbase Exchange format. This standard allows for the interchange of terminology data including detailed lexical information. The framework for TBX is provided by two ISO 12620, ISO 12200 and ISO Committee Draft 16642, known as TMF or Terminological Markup Framework. ISO 12620 provides an inventory of well-defined "data categories" with standardized names that function as data element types or as predefined values.
- **OLIF:** Open Lexicon Interchange Format. OLIF is an open, XML-compliant standard for the exchange of terminological and lexical data. Although originally intended as a means for the exchange of lexical data between proprietary machine translation lexicons, it has evolved into a more general standard for terminology exchange.

Automatic Acquisition

UNIVERSITÄT DES SAARLANDES

- Technical terminology exists in large and quickly growing quantities
- A good way of keeping track is to mine existing documents for technical terms
- Statistical criteria can be used for monolingual term extraction, but based on frequencies alone it is hard to separate the wheat from the chaff
- Recent effort at DFKI (in collaboration with European Patent Office): Automatic terminology extraction from translated documents

Terminology Extraction from Patents

UNIVERSITÄT DES SAARLANDES

Cooperation between DFKI and European Patent Office (EPO)

- Goal: Extract parallel terminologies for EN, DE, ES, FR from translated patent documents
- Motivation:
  - Technical documentation makes up a large share of language industry's raw material, vocabulary is commercially interesting
  - Manual construction of unrestricted multilingual terminologies would be prohibitively expensive
  - Translated documents exist in large volumes, as well as techniques for sentence/word/phrase alignment
  - IPC (hierarchical system of about 70K classes) may help to relate extracted terms with ontologies
  - Test-bed for scalability of tools and resources
    - How well do our tools cover technical texts?
    - Can we acquire new lexical information from data?
  - First step towards MT for technical documents

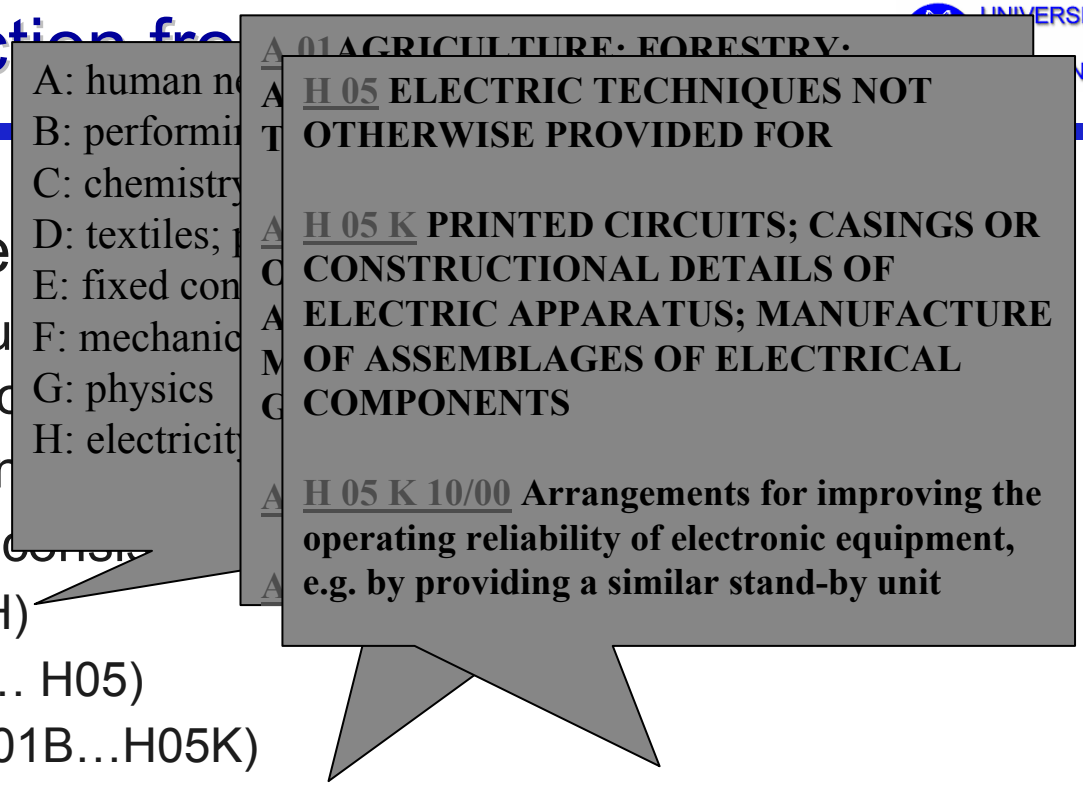
Terminology Extraction from Patents

UNIVERSITÄT DES SAARLANDES

History and current status:

- Techniques were prototypically implemented in a feasibility study for WIPO (‘03, via acrolinx GmbH)
- Call for Tender by EPO in August ‘05
- Bids and results on test data due in September
- EPO received 14 bids, DFKI delivered best results for DE↔EN, ES↔EN and was among the best for FR↔EN
- System for production under construction, started processing first batch of data (2.9M docs, >90GB)

Terminology Extraction fro



The International Pate

- based on the Strasbou
  used by >100 natic
- indispensable for findi
- hierarchical structure, c
  - eight sections (A..H)
  - 120 classes (A01 … H05)
  - 628 subclasses (A01B…H05K)
  - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 7th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities

Terminology Extraction from Patents

UNIVERSITÄT DES SAARLANDES

Some research questions related to the IPC

- Automatic Classification
  - Can IPC classes be identified automatically?
  - *(So far classification and search done by ~ 6500 experts)*
- Ontology construction
  - How does the IPC relate to the terminologies used in the various domains? Can we (semi-) automatically construct/extend these terminologies given the documents?
- Word sense disambiguation
  - Can a given IPC class help to identify meaning/translation of a given term?

Terminology Extraction from Patents

UNIVERSITÄT DES SAARLANDES

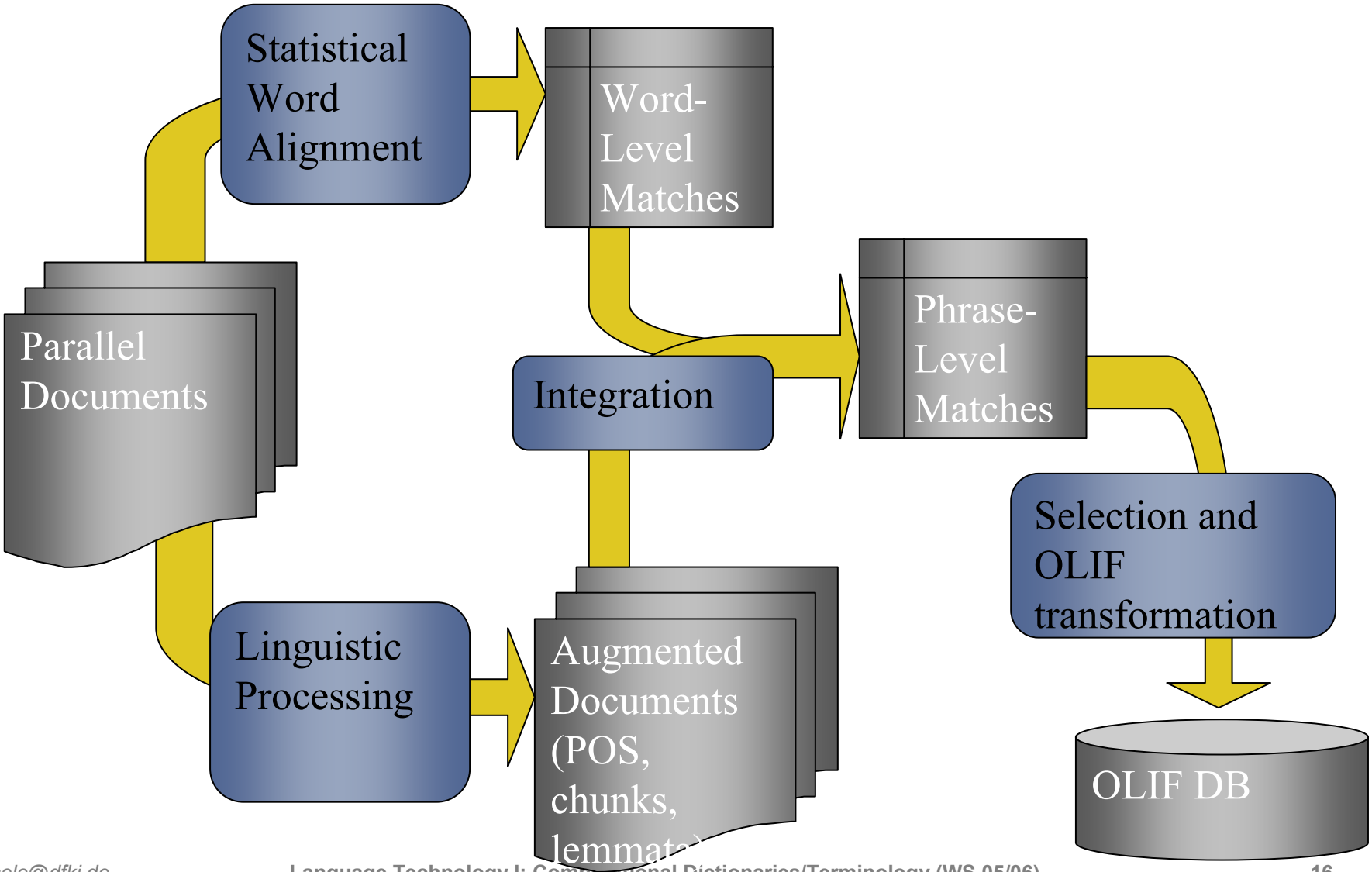
Setup:

- Use linguistic tools for corpus annotation
  - POS-tagging
  - Phrase recognition
  - Lemmatization
- Use statistical tools for alignment
  - GIZA++ from Franz Och
  - Own algorithms based on word similarities
- Integrate module outcomes, transform into OLIF entries

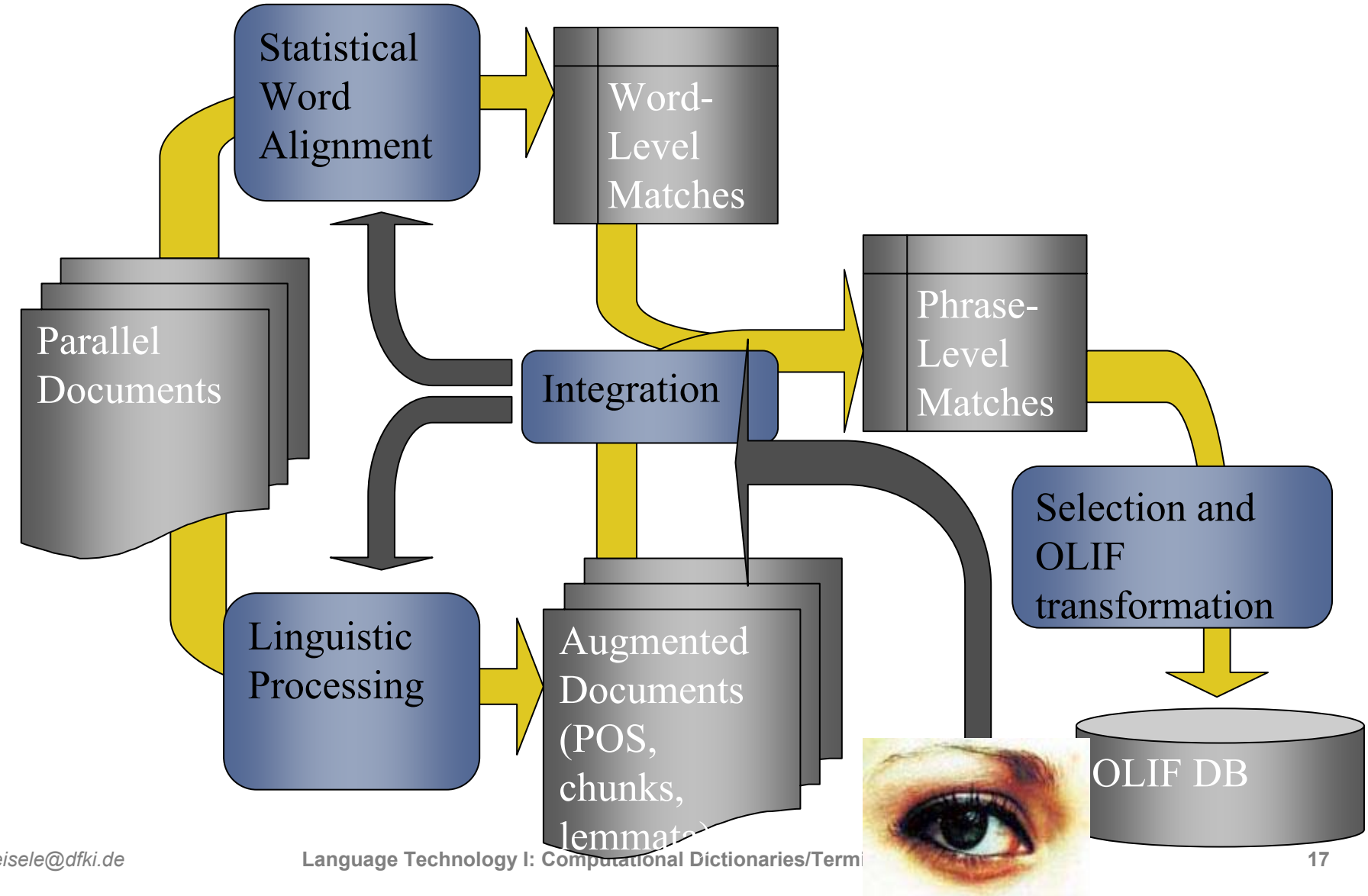
Improvement in 2nd phase:

- Feed-back of modifications to basic modules
- Manual inspection and error analysis will be used to improve algorithms as long as the project is ongoing

Terminology Extraction: Architecture



Terminology Extraction: Architecture II



Examples for Patent Terminology

UNIVERSITÄT DES SAARLANDES

Postbestimmungsortinformationsspeichereinrichtung
= mail destination information memory means

Informationsdurchforstungssteuerungseinrichtung
= information browsing control means

Hypervideonachrichtversendungsverarbeitungseinrichtung
= hypervideo message posting processing means

Gasphasenverunreinigungsabsorptionsflüssigkeit
= gas phase contaminant absorbing liquid

Exercises

UNIVERSITÄT DES SAARLANDES

- Do a web search for a freely downloadable English-German bilingual lexical resource (try to find a large one). How many entries does it contain? How many translations for the words “aktuell” and “offender” can you find in it? Which of these terms can you find in wordnet?
- Compare these numbers with the ones given in a recent (superficial) comparison of web-based dictionaries in http://tomorrow.msn.de/internet/webguides/web-dolmetscher
- Look through a list “translations” of aktuell (+inflection) extracted by a statistical approach from a parallel corpus. Are there good entries in the automatically found data the lexicon was missing?
- Look in the ELRA catalog for English-German bilingual lexical/ terminological resources. Can you find anything useful? How much would it cost? Is it possible to obtain information about the level of detail contained in these resources without buying them?