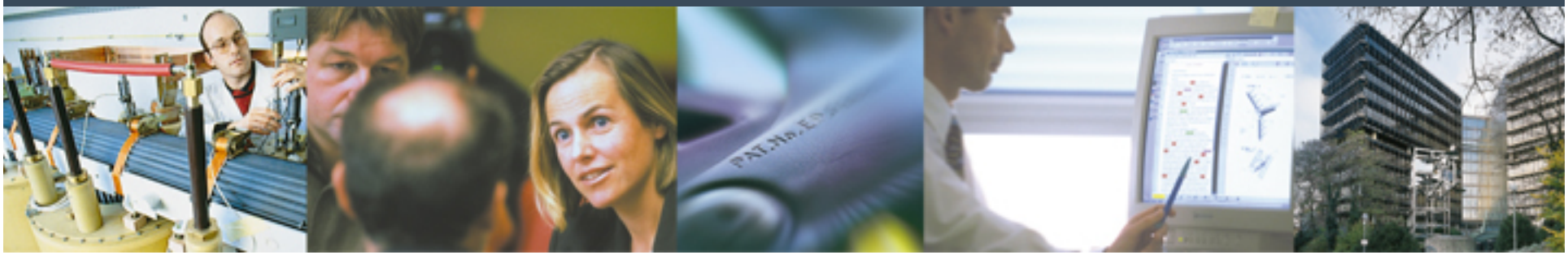
European
Patent Office

European Machine Translation Programme

Wolfgang Täger

December 2006



European
Patent Office

Overview

- Programme Partners and Goals
- MT engine
- Dictionary format
- Available corpora
- Alignment & Extraction
- Validation & Concordancing

- **DEMO**

European Patent Office

Programme Partners and Goals

- Trigger: Success of JP-EN patent translation
- Agreement EPO - Member States
  1. MT of patents/ abstracts/ communications to/from English
  2. Three language pairs per year
  3. First three languages: FR - DE - ES
- Candidates for next year: Swedish, Dutch, Italian, Romanian, Greek

European
Patent Office

MT engine

Trial with SMT system (Language Weaver)

Call for tender: Winner Worldlingo (Systran)

Going public (esp@cenet): December 2006

Needed: Improve translation by specific dictionaries

European
Patent Office

Dictionary format

Desiderata

- open standard
- XML-Unicode
- support features of MT engines
- support conditional translations (e.g. based on IPC)

Is not intended for terminology (no definitions, lexical focus and no semantic focus).

⇒OLIF format was chosen

How to get dictionaries ? By bilingual term extraction !

European
Patent Office

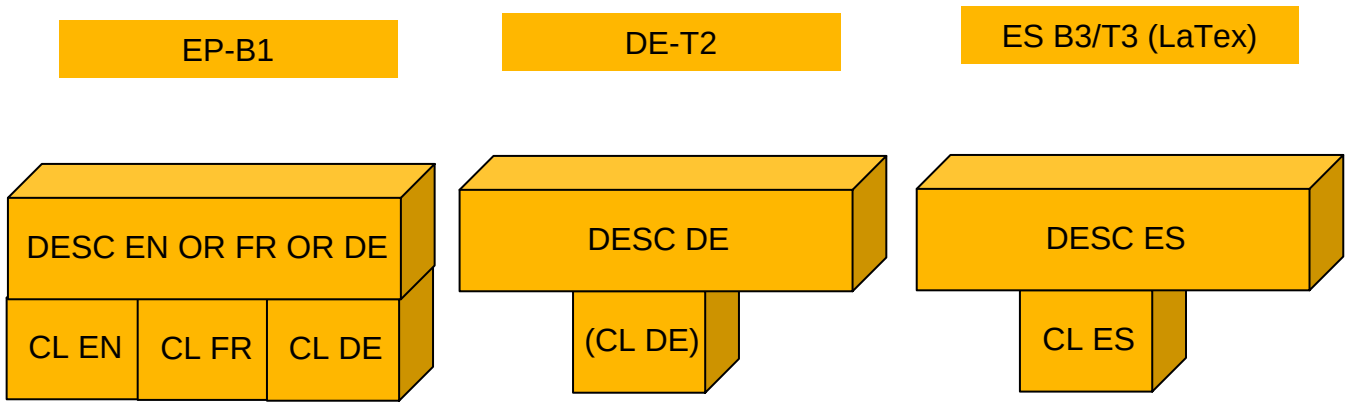
Available corpora

560.000 EP-B publications => claims in EN,DE,FR

300.000 DE-T2 publications

37.000 ES-B3/T3 publications

=> Align corpora for term extraction, concordancing, translation memory (and SMT)



European
Patent Office

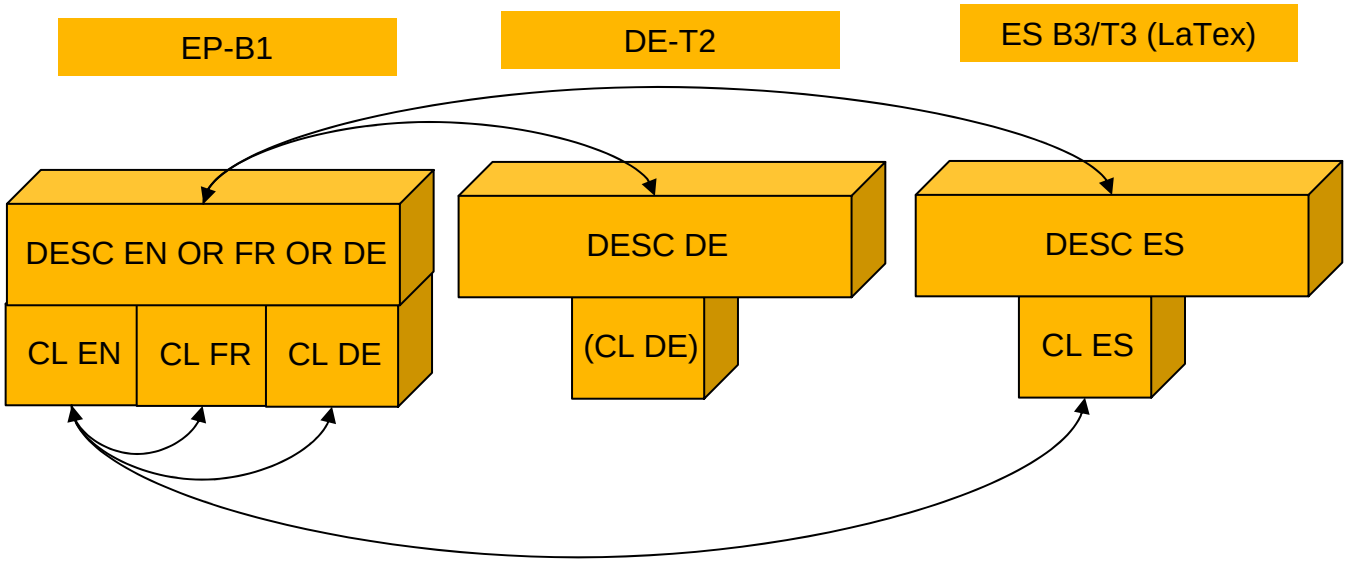
Available corpora

560.000 EP-B publications => claims in EN,DE,FR

300.000 DE-T2 publications

37.000 ES-B3/T3 publications

=> Align corpora for term extraction, concordancing, translation memory (and SMT)



European
Patent Office

Alignment & Extraction

Alignment: Trial at EPO with internally developed SW

Result was not improved by external companies during call for tender.

European
Patent Office

Alignment & Extraction

Call for tender for bilingual term extraction

Winner: DFKI

1. Alignment of corpora, POS tagging, Identification of terms
2. Pairing of terms using clues like co-occurrence score, string similarity, grammatical clues, position, available dictionaries, ...
3. Providing further information like gender, inflection, transitivity, countable, ...

European
Patent Office

Validation & Concordancing

Development of OLIF editor at EPO

- Remove noise
- Correct entries
- Use concordancer (provides statistics based on parallel corpora)

=> DEMO

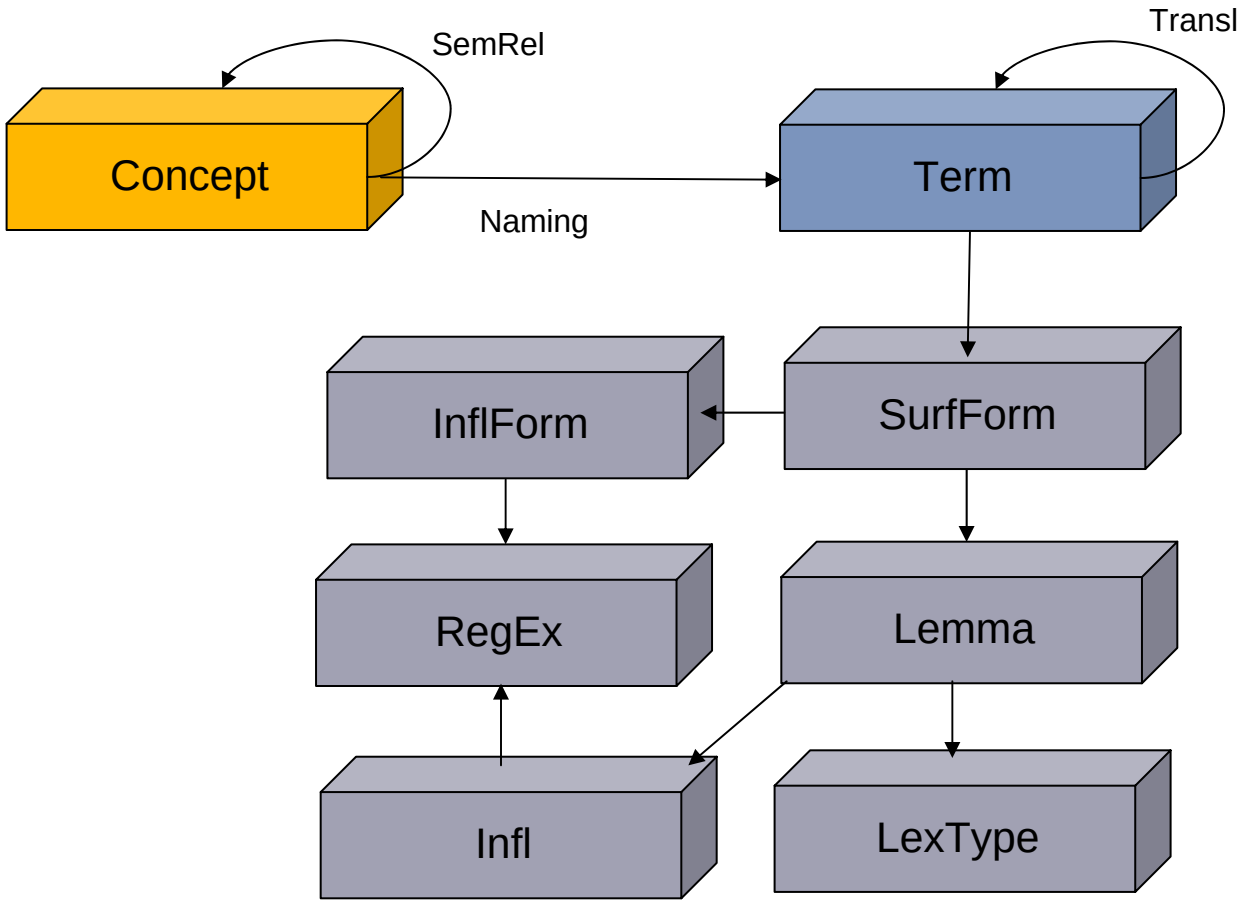
European
Patent Office

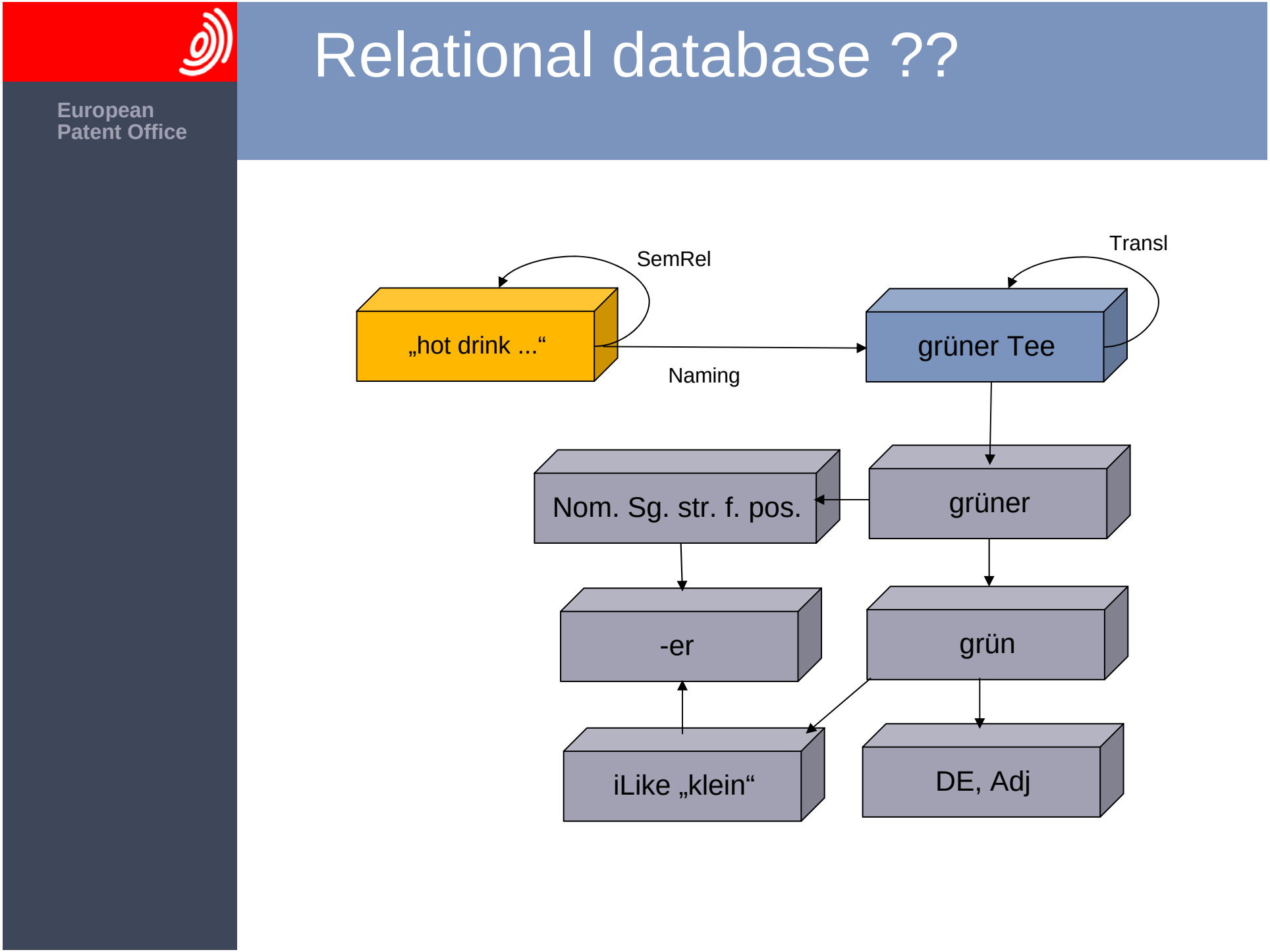
OLIF format

- Support of more languages
- Clarification of inflection scheme
- Clarification of term vs lex approach
- Tools

European
Patent Office

Relational database ??





European
Patent Office

Relational database ??

European
Patent Office

End

Thank you!