OLIF V2

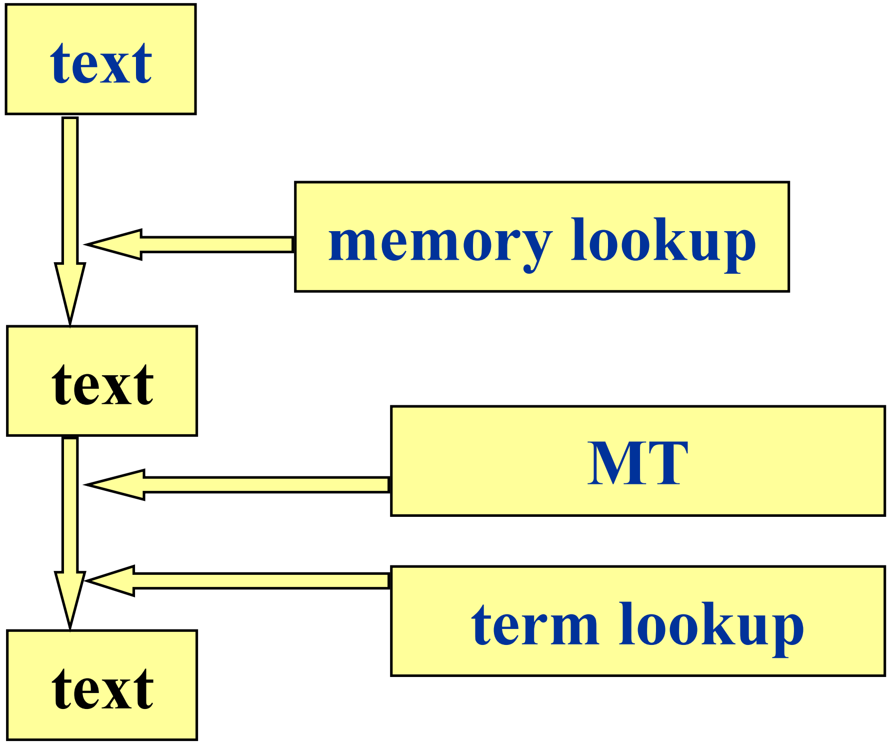
SAILLABS

Gr. Thurmair
April 2000

OLIF: Overview

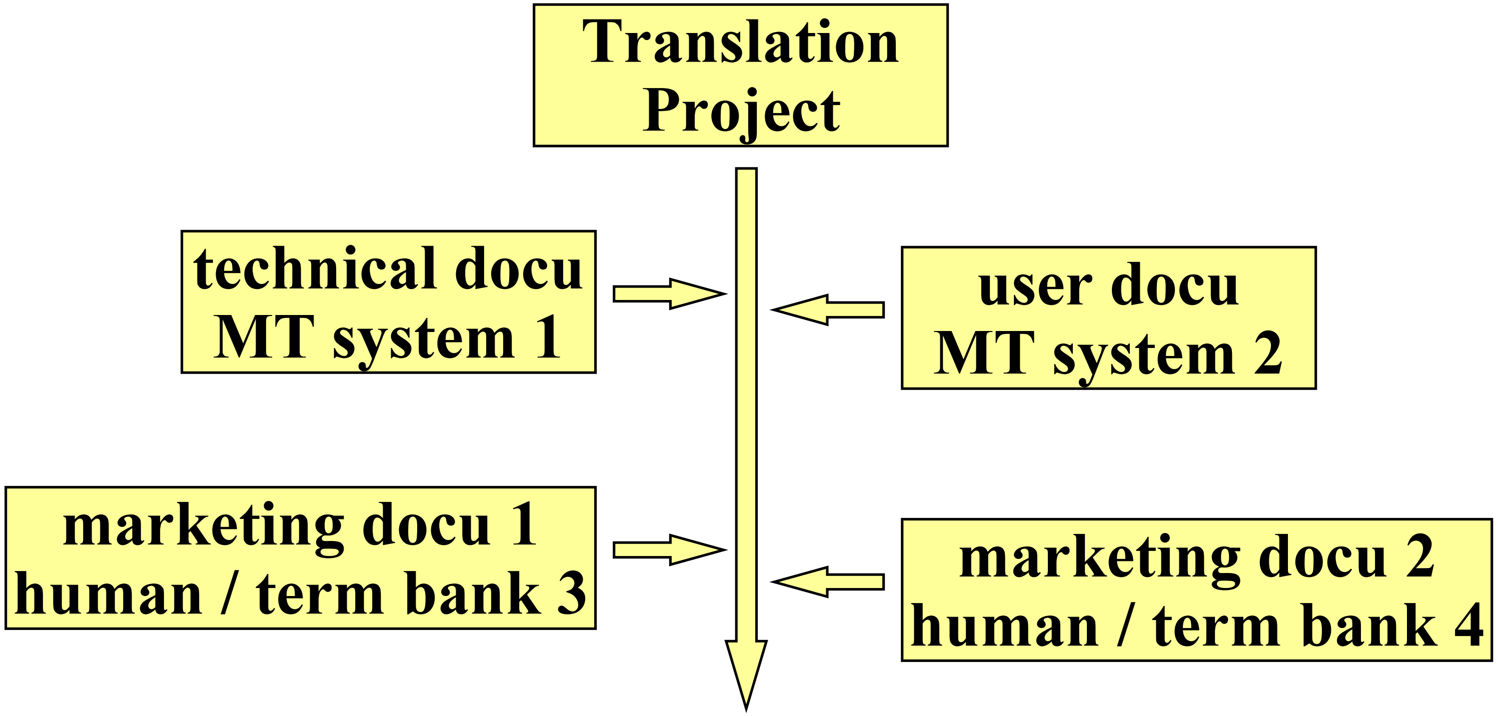
- Rationale
- Principles
- Entries
- Descriptions
- Header
- Examples
- Status

The Exchange Problem



- the same text is processed with different tools
- how do these tools communicate?

The Resource Problem



Translation Project

technical docu MT system 1

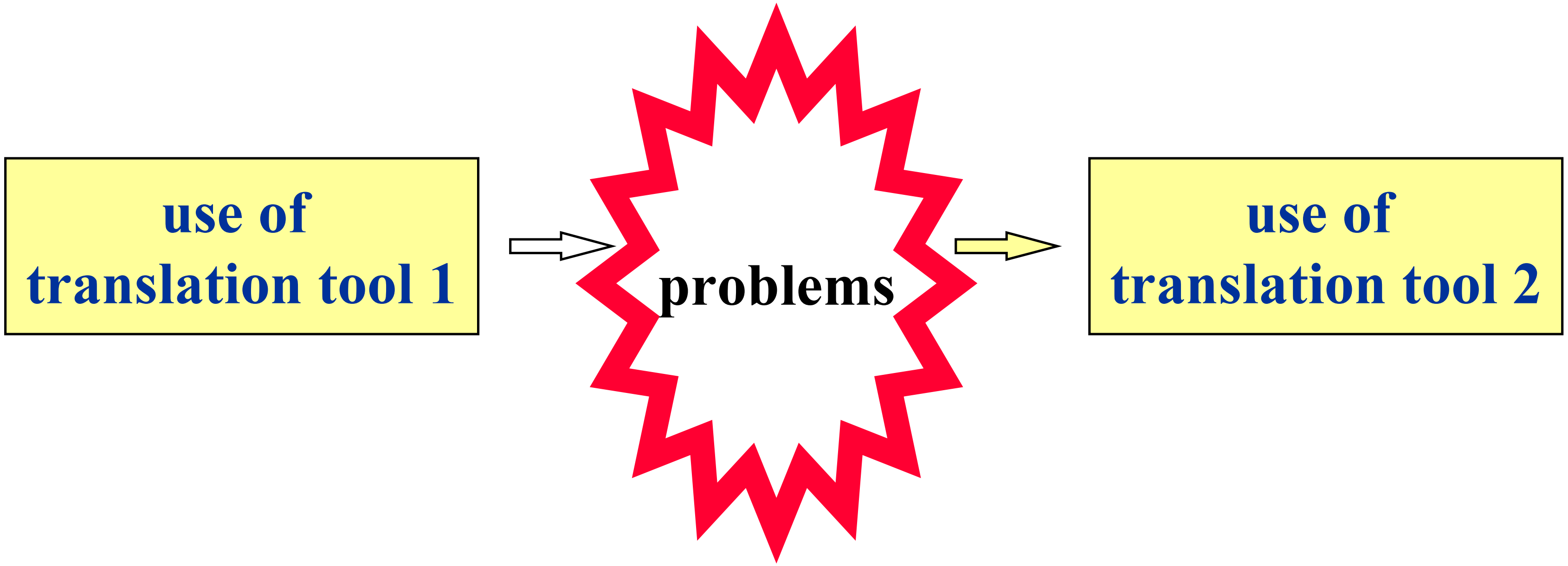
user docu MT system 2

marketing docu 1 human / term bank 3

marketing docu 2 human / term bank 4

- how often will the same term be stored?
- who pays for redundant maintenance?

The Migration Problem



use of translation tool 1

problems

use of translation tool 2

- do you have to re-build your language resources?
- who pays for rebuilding lexicons and memories?

Problems in Exchange

- Different *purposes* of exchange
  - data import / export
  - data validation
- Different *content* of exchange
  - terminological data, different for each system
  - lexicographical data, different for each system
- Different *structures* of exchange files
  - trend towards markup structures (SGML/XML)

SAILLABS

OLIF: Principles

- Keep it simple!
  - flat feature value structures
  - standard software environment
- Keep it pragmatic!
  - worry only about what’s there
  - bottom-up: compare what systems *have*

OLIF V2 formalism

- Define a representation formalism:
  - XML
    - general format; processing tools available
    - but: overhead in markups
      => not suitable for *very large* files
    - Coding standard: UTF-8
  - File structure
    - header: globals, defaults, definitions
    - body: sequence of entries

OLIF principles

- Entries are *concept*-based
  - i.e. we describe *word senses:* different readings -> different entries
- Entries have (monolingual) *descriptions*
  - both for MT and terminology
- Entries have *links* to other entries
  - inner-language: crossreference links
  - intra-language: transfers (multilingual, directed)

SAILLABS

Entry Structure

- “central” information
  - definition features
  - administrative features
- monolingual / linguistic feature set
- terminological feature set
- transfer features
- cross-references / links

Definition of the entry

- Obligatory features:
  - language
    - (only language or also locale?)
  - canonical form
    - do we need guidelines (multiwords)
  - part of speech
    - only open word classes: N, V, A, Adv, Prep
  - domain information (semantics)
    - => a common top level classification?
  - reading no.

Scope of the descriptions

- minimal linguistic descriptions
  - features which everybody has / needs
- minimal terminology descriptions
  - feature set of e.g. Interval
- minimal transfer descriptions
  - equitype, tests, transfers
- minimal thesaurus / ontology relations
  - ISO standards
- additional fields for “personal” use

Linguistic Descriptions

- Morphological Features
  - entry type
    - abbreviation, single word, compound, multiword)
  - inflection class
    - enumeration of inflection patterns, per category
  - gender
  - (special) number
    - singulare / plurale tantum
  - degree / comparative

SAILLABS

Linguistic descriptions

- Syntactic Features
  - Syntactic Type
    - (subcategorisation of part-of-speech)
  - Syntactic Frame
    - Argument structures (DObj, PObj-for, ...)
  - (Transitivity)
    - intransitive, transitive

SAILLABS

Linguistic Descriptions

- Semantic Features
  - Semantic type
    - for subclasses only?
    - who has / needs it?

Terminological Descriptions

- Minimum needed to validate an entry (Interval)
  - Definition
  - Context
  - Scope
  - Comment / Note
  - (validation status)
    - (a three-level hierarchy)

SAILLABS

Transfer Descriptions

- Equivalence type
  - full - partial (subset / superset) - none
  - for reversible entries
- Tests and Actions
  - (to be worked out)
- Comment
- (Definition of the “target” link)

Cross-Reference Descriptions

- Linktype
  - thesaurus relations
    - broader / narrower / synonym / related
  - additional customisable relations
    - abbreviation_for, forbidden, outdated
- Definition of the “target” link

Administrative Information

- Source of Entry
  - string
- Author
  - creation author
  - last modofication
- Date
  - creation date
  - last modification date

Header Information

- Definition of Encoding
  - given in the XML statement (UTF-8)
- Definition of features / values used
- Definition of default values

Example (1)

```
<ENTRY>
          <MONO>

                   <LG>    de     </LG>
                   <CAN>  Brot    </CAN>
                   <CAT>  noun    </CAT>
                    <SA>  gv      </SA>
        ....
           </MONO>
 </ENTRY>
```

SAILLABS

Example 2a

```
<Entry>
<MONO>
....
  <CAN>    offshore account </CAN>
  <CAT>    noun </CAT>
  <SA>     Money-Laundering </SA>
 ...
</MONO>
<XFR>
  <CAN>   compte en banque à l‘étranger </CAN>
  <LG>     fr </LG>
  <CAT>   noun </CAT>
  <SA>     Money-Laundering </SA>
</XFR>
<XFR>
  <CAN>  Auslandskonto </CAN>
  <LG>     de </LG>
...
</Entry>
```

Example (2b)

```
<Entry>
<MONO>
   <CAN> compte en banque à l‘étranger </CAN>
   <CAT>noun </CAT>
   <SA>Money-Laundering </SA>
...
</MONO>
<XFR>
   <CAN>offshore account </CAN>
   <LG>en </LG>
   <CAT>noun
   <SA>Money-Laundering </SA>
</XFR>
<XFR>
<CAN>Auslandskonto </CAN>
<LG>de </LG>
...
</Entry>
```

OLIF: Status

- Implementation of a central DB
- Implementation of OLIF parser & generator

BUT:

- different “flavours” of OLIF
  - dependent on projects (OTELO - Aventinus)
  - different formalisms (own, SGML, XML)

SAILLABS

Status

- Verification of OLIF
  - MT lexicons (Logos, T1)
  - Term Bases (SAPterm, DanTerm)
- Converter prototypes
  - T1 <-> OLIF, Logos <-> OLIF
- Term Lookup systems
  - based on OLIF-type database
- Comparisons with other formats