Language Resources, Language Technology, Text Mining, the Semantic Web: How interoperability of machines can help humans in the multilingual web

Felix Sasaki
DFKI / University of Appl. Sciences Potsdam
W3C German-Austrian Office
felix.sasaki@dfki.de

Purpose of this talk (1)

- Show gaps
  - Between machines
  - Between machines and humans
- … which we need to fill to bridge gaps between humans

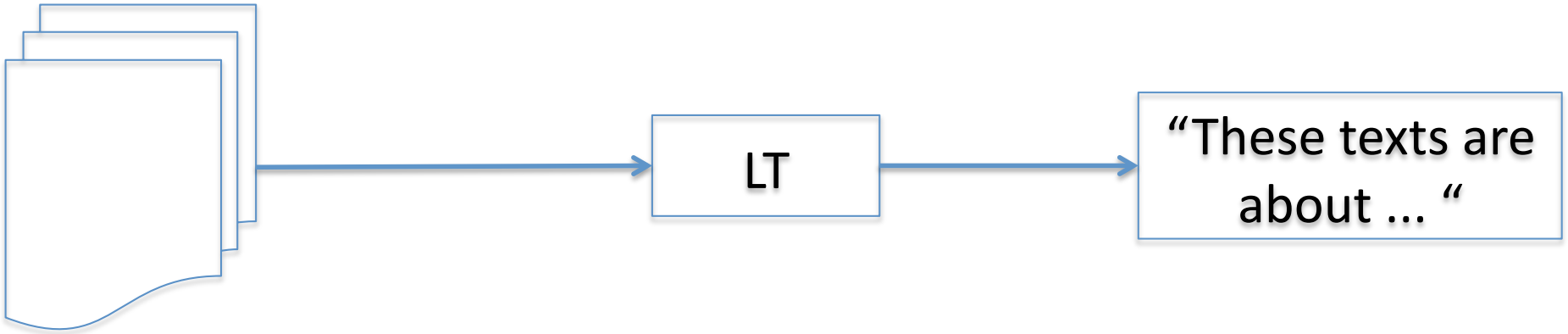
Purpose of this talk (2)

- Identify groups / communities
  - To fill gaps
  - To come together in new alliances

Basics:
What are machines doing (not only on the Web)?

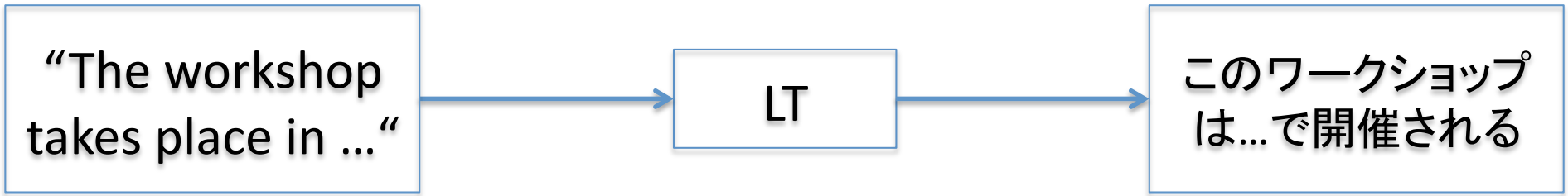
Language Technology

- Summarization



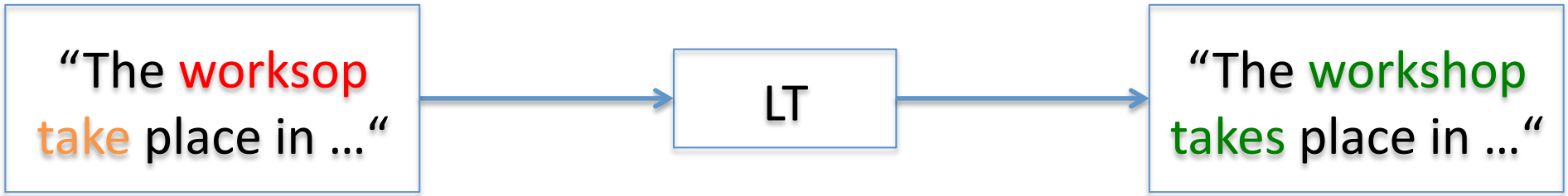
Language Technology

- Machine Translation



Language Technology

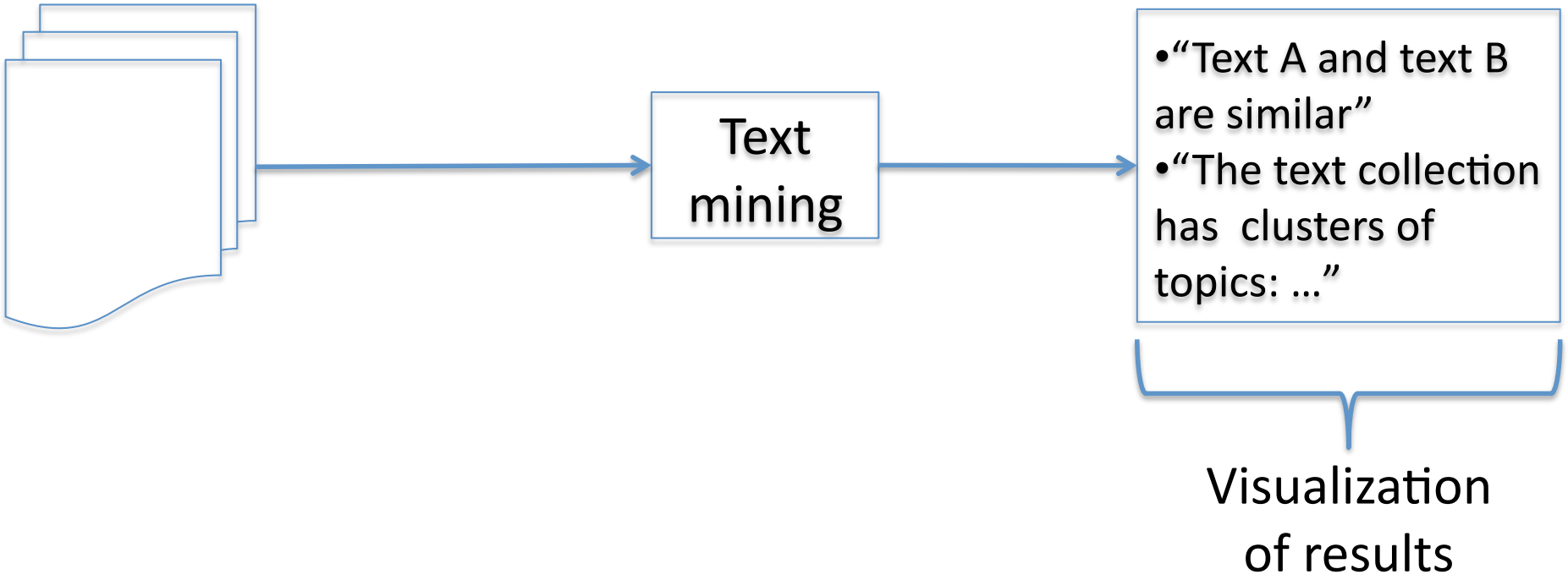
- Spell and grammar checking



- And many more applications
  - Coreference resolution, discourse analysis, named entity recognition, natural language generation, question answering, …

Text mining

- Finding out things you did not know



Basics:

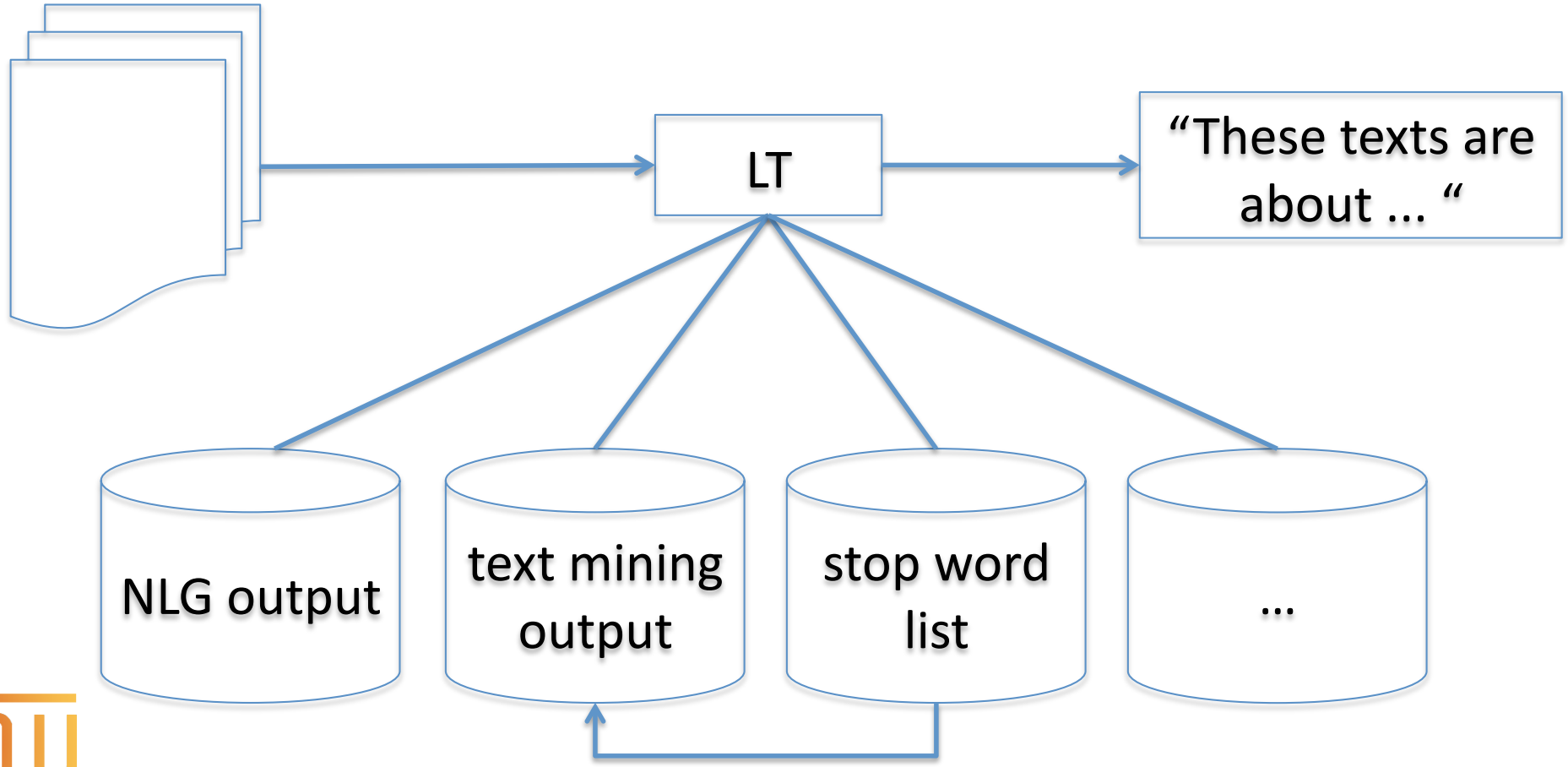
What are machines doing (not only on the Web)?

How are they doing it?

They are using resources

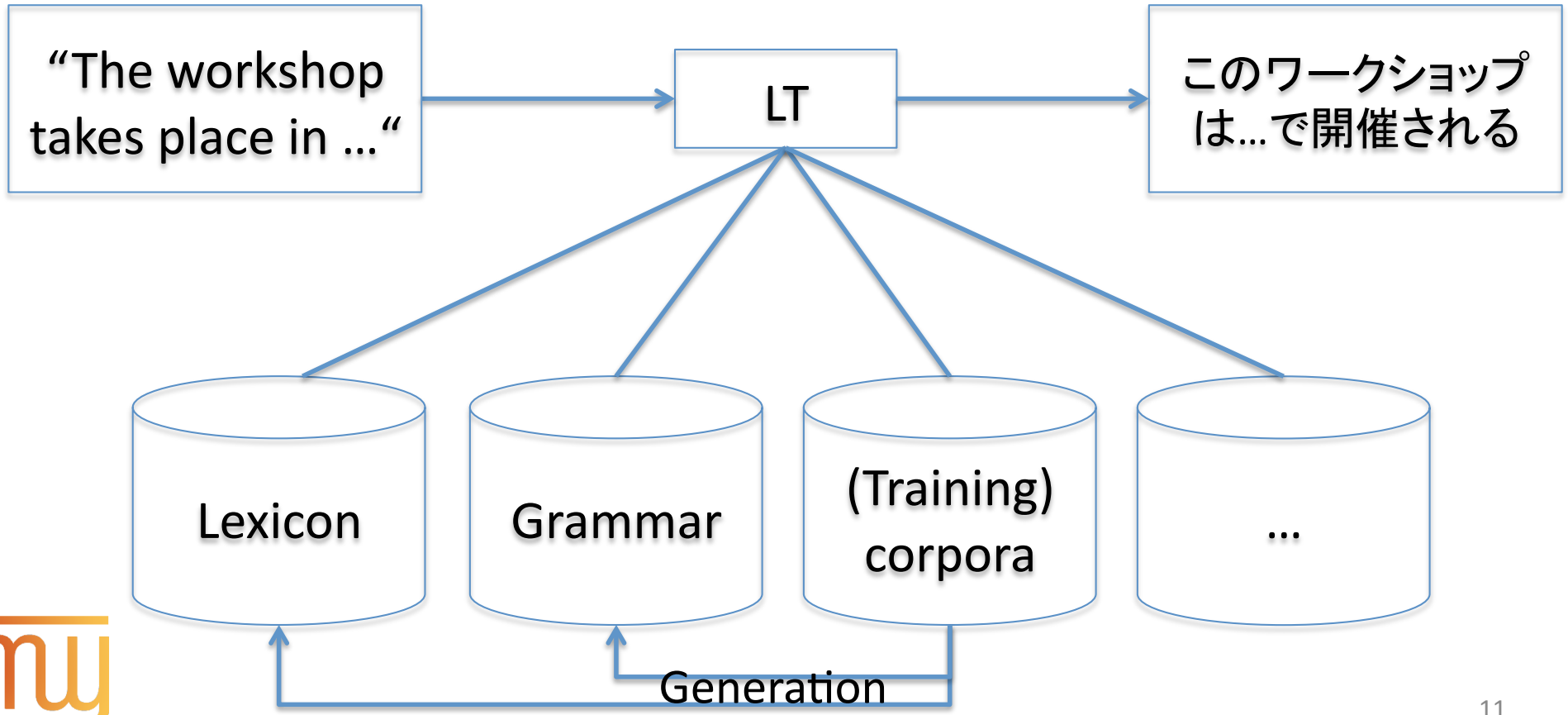
Resources in language technology

- Sample resources for summarization



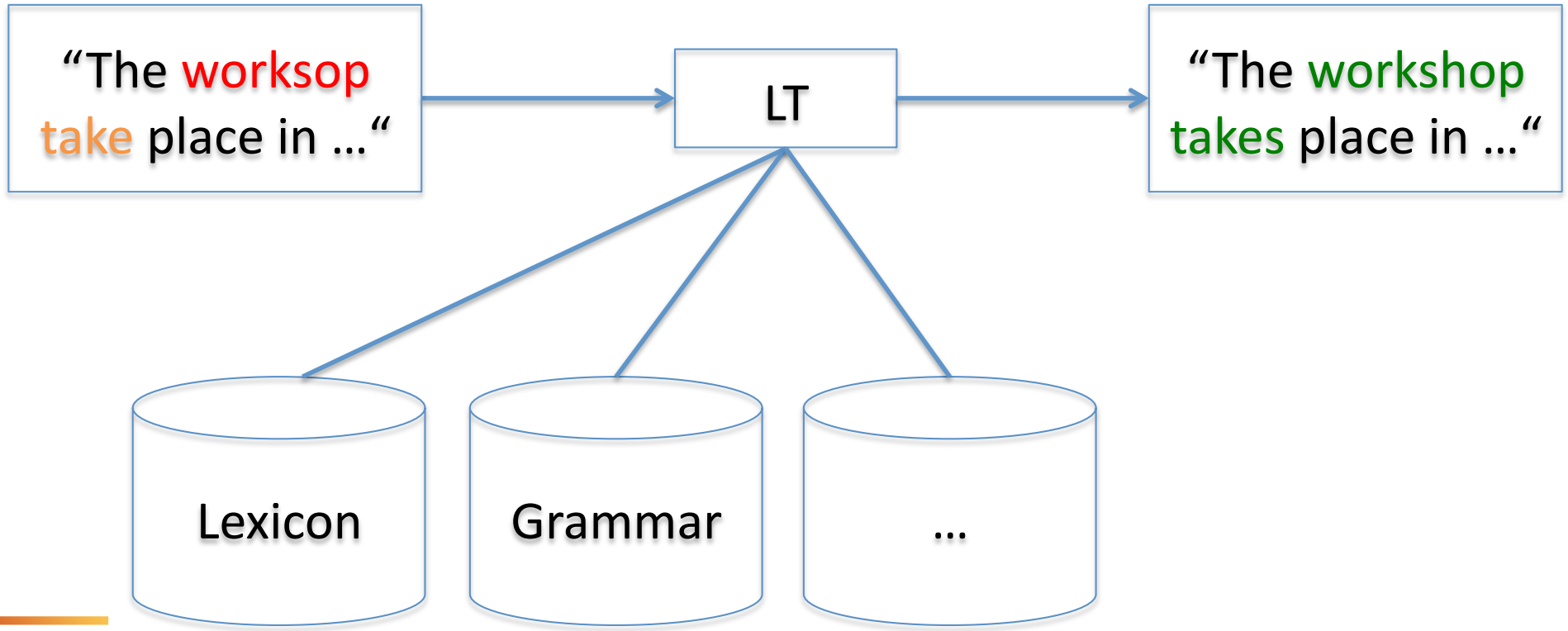
Language Technology

- Sample resources in Machine Translation



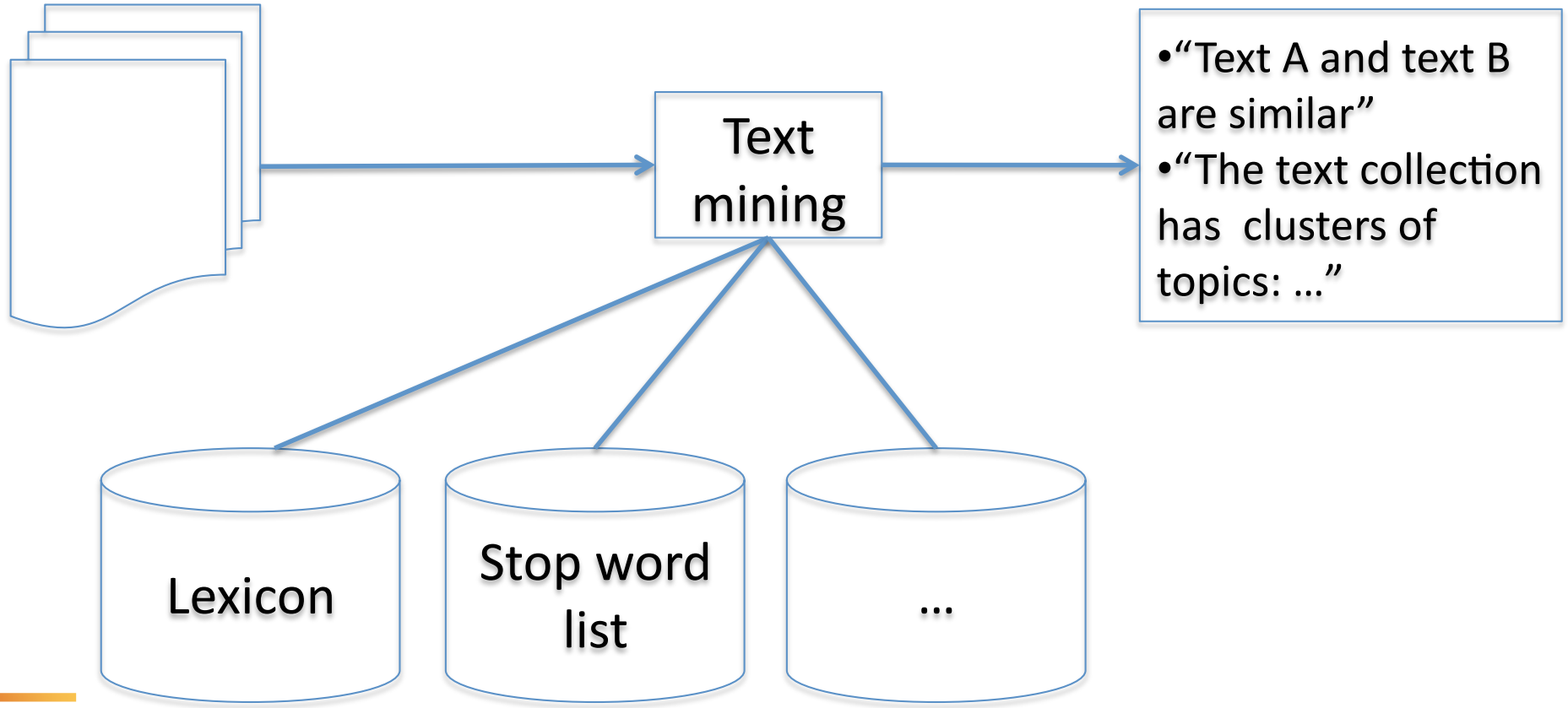
Language Technology

- Sample resources for spell and grammar checking

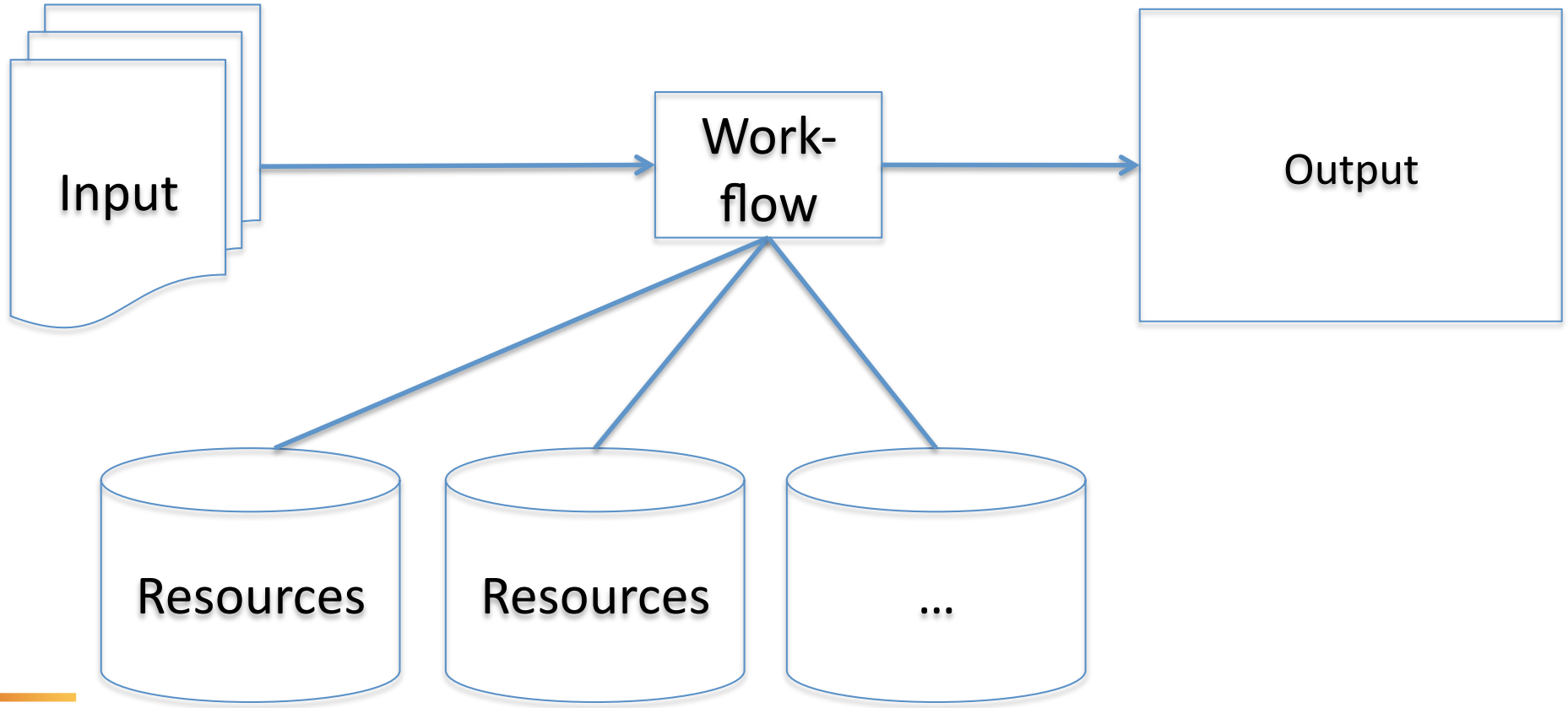


Text mining

- Sample resources for text mining



In general: you need three types of data: input, resources, workflow



What gaps need to be filled for truly "multilingual content processing"?

- Gap 1: machines don't use metadata available in the input
- Gap 2: machines don't know about the workflow (input) data goes through
- Gap 3: machines don't make explicit
  - "Who" they are
  - What resources they are using

Gap 1: machines don't use metadata available in the input

- Input from www.postbank.de

  „Ob Postbank direkt, Online-Banking, Online-Brokerage oder myBHW. Die häufigsten Fragen zu unseren Transaktionssystemen finden Sie an dieser Stelle."

- Output via Google translate

  "Whether Postbank direct, online banking, online brokerage or myBHW. Frequently asked questions about our transaction systems can be found at this location."

Gap 1: machines don't use metadata available in the input

- Input from www.postbank.de

  „Ob Postbank direkt, Online-Banking, Online-Brokerage oder myBHW. Die häufigsten Fragen zu unseren Transaktionssystemen finden Sie an dieser Stelle."
- Output via Google translate

  "Whether Postbank direct, online banking, online brokerage or myBHW. Frequently asked questions about our transaction systems can be found at this location."

Fixed terminology should not have been translated.

But – the MT tool had no chance to "know" that – why?

Gap 2: machines don't know about processes data goes through

- Input from the data base – the "hidden web":

  „Ob <term>Postbank direkt</term>,
  <term>Online-Banking</term>,
  <term>Online-Brokerage</term> …"

- Output on the Web:

  „Ob Postbank direkt,
  Online-Banking,
  Online-Brokerage …"

fixed terminology (= metadata) …

publication process

… is lost on the Web ☹

Gap 3: no common identification …

- Of metadata and processes chains (previous slides)
- Of resources – e.g. what is a lexicon
  - In machine translation?
  - In localization?
  - For a human reader?
  - Ability to combine tools depends on knowing about them (capabilities, resources) in detail

Who can fill these gaps – people dealing with multilingual content

- Content producers
  - Allow for terminology identification in source formats / CMS
- Localizers
  - Make localization workflows aware of (process / source content) metadata
- "Machine" experts
  - Make their tools sensible to source content metadata and expose their capabilities (what resources / workflows) in a clear defined way

Who can fill these gaps – people dealing with multilingual content

- Users
  - Add metadata to source content
  - Use (machine translation) tools without knowing the details – e.g. in the browser!
- Browser vendors
  - Create APIs which make use of automatic tools / resource and workflow descriptions / source code metadata
- …

➔ The people in this room!

How can they fill the gaps?

- All these groups need to agree upon one machine readable information space for filling the gaps
- It's actually already here – the Semantic Web!

What is the Semantic Web

- The Web as humans see it: Identification of "meaning" e.g. via (*typographic* or other) conventions

„Ob *Postbank direkt* …"

What is the Semantic Web

- The Web as machines see it: Identification of meaning via RDF-based mechanisms (here via RDFa)

„Ob Postbank direkt …“

What is the Semantic Web – RDF in 30 seconds

- A framework for making statements about resources, using URIs
- RDF can help to fill our gaps
  1. Metadata in the input
  2. Metadata for workflows
  3. Identify 1., 2. and language technology resources uniquely
- In one information space – the machine readable Web

Instead of a summary – call for project (participating in) proposals

- Who needs to come together
  - Content producers, localizers, "machine" experts, browser vendors, users
- What should their work be based upon
  - Semantic Web technologies
  - Clear interfaces to the human (e.g. browser) Web, like RDFa
- What we do not need
  - Web-centred standardization of formats for language resources themselves – that is already done elsewhere (see this session)
- Where the place is to do that work?
  - W3C, since it needs to be part of core Web technologies
- For making it happen, we need a strong alliance of Web technologies, other fields and machine technologies

META-NET

- EU-funded project, closely related to "Multilingual Web"
- Main aim: build an alliance for improving language technologies in Europe
- Laaarge: soon 40+ participating organizations in 30+ countries
- Very important: bring users of language technology in

META-NET

- Users and language technology companies = in Europe not only large companies, but more and more small SMEs
- Target of META-NET are these small and fast units – including you ☺
- EU has started special funding programs for SMEs – see http://tinyurl.com/eu-lt-sme ("objective 4.1")

META-NET

- Event: META-NET Forum
- Brussels, November 17th/18th
- Aim: Bring users / language technology developers / policy makers together
- Discuss a road map for the next 10 years of language technology road map and its applications
- Details and registration at

http://www.meta-net.eu/events

Language Resources, Language Technology, Text Mining, the Semantic Web: How interoperability of machines can help humans in the multilingual web

Felix Sasaki
DFKI / University of Appl. Sciences Potsdam
W3C German-Austrian Office
felix.sasaki@dfki.de