Computational Linguistics in the Industry

Authoring Support with acrolinx IQ ™

acrolinx
TECHNOLOGY FOR INFORMATION QUALITY

overview

- acrolinx – the company
- production of technical documents
- NLP for
    - spelling and terminology
    - grammar
    - style
    - consistent phrasing

acrolinx – the company

- software for information quality assurance
- spin-off from German Research Center for Artificial Intelligence (DFKI), Saarbrücken
- technology under development since 1997 (since 2002 as acrolinx))
- headquarter in Berlin, about 40 employees
- users in 25 countries, checking millions of words a month

some of our customers

Software	Life Sciences	Communications	Industrial	Technology
Adobe	Dräger	AlcatelLucent	DAF	Bosch
Autodesk	GE	Cisco	HOMAG	Embraer
CA	Medtronic	Huawei	John Deere	KonicaMinolta
EMC	Siemens	Motorola	MAN	Philips
IBM		SonyEricsson	SEW Eurodrive	
SAS Institute			Siemens	
Symantec			Leica GeoSystems	

production of technical documents

- correctness
- understandability
- readability
- translatability
- consistence
- less ambiguity
- corporate wording

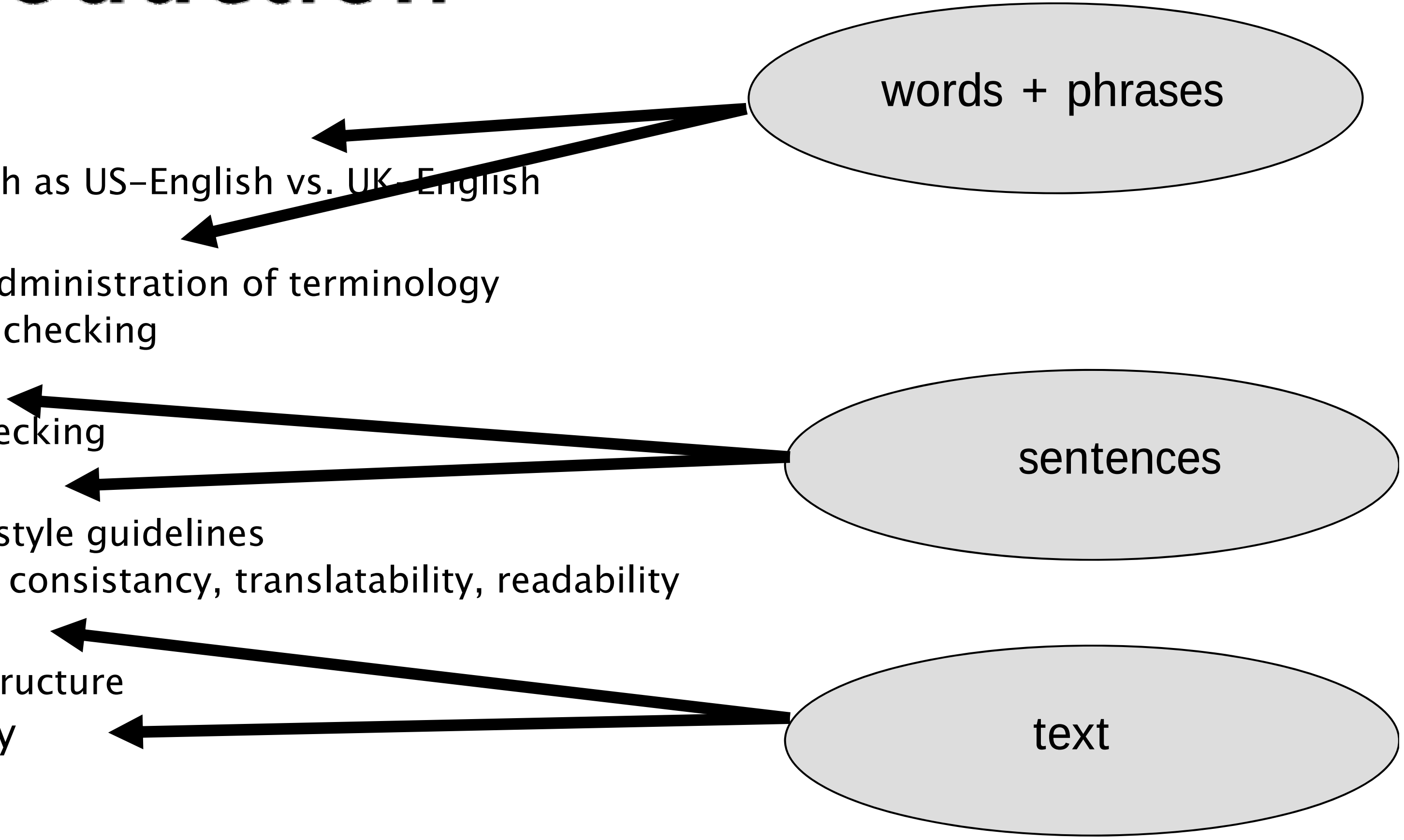
- spelling
- grammar
- style
- terminology

QA is a Cost Factor

- Translation costs
- Support costs

text production

- spelling
  - variants, such as US-English vs. UK-English
- terminology
  - set up and administration of terminology
  - terminology checking
- grammar
  - grammar checking
- style
  - checking of style guidelines
  - checking for consistancy, translatability, readability
- structure
  - document structure
- multilinguality



words + phrases

sentences

text

spelling

- words are defined in a dictionary
- anything not in the dictionary is an error
- high recall, low precision (depending on the domain)

language analysis

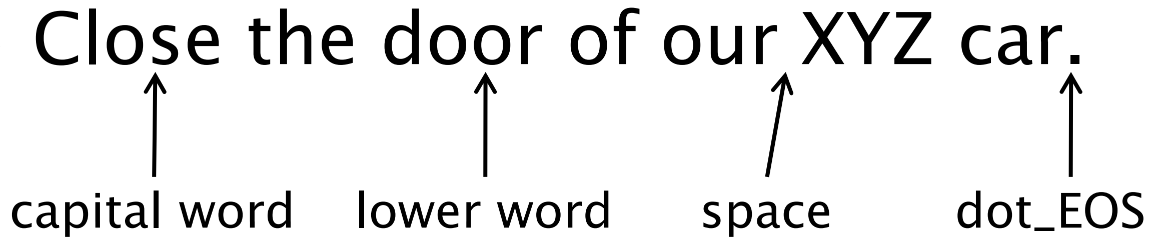
- errors are defined
- unknown words that are not defined as errors are term candidates
- based on words and rules
- consider terminology
- high precision, recall is dependent on data work

error analysis

NLP for words and phrases

- tokenization
- POS-tagging
- morphology
- dictionary
- error dictionary

tokenization



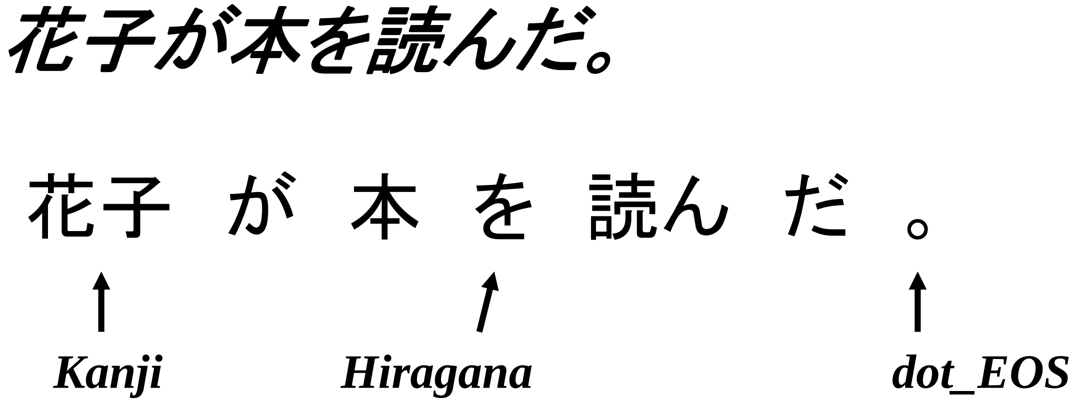
- Close the door of our XYZ car.

capital word

lower word

space

dot_EOS



花子が本を読んだ。

花子 が 本 を 読ん だ 。

Kanji

Hiragana

dot_EOS

based on rules and lists of abbreviations

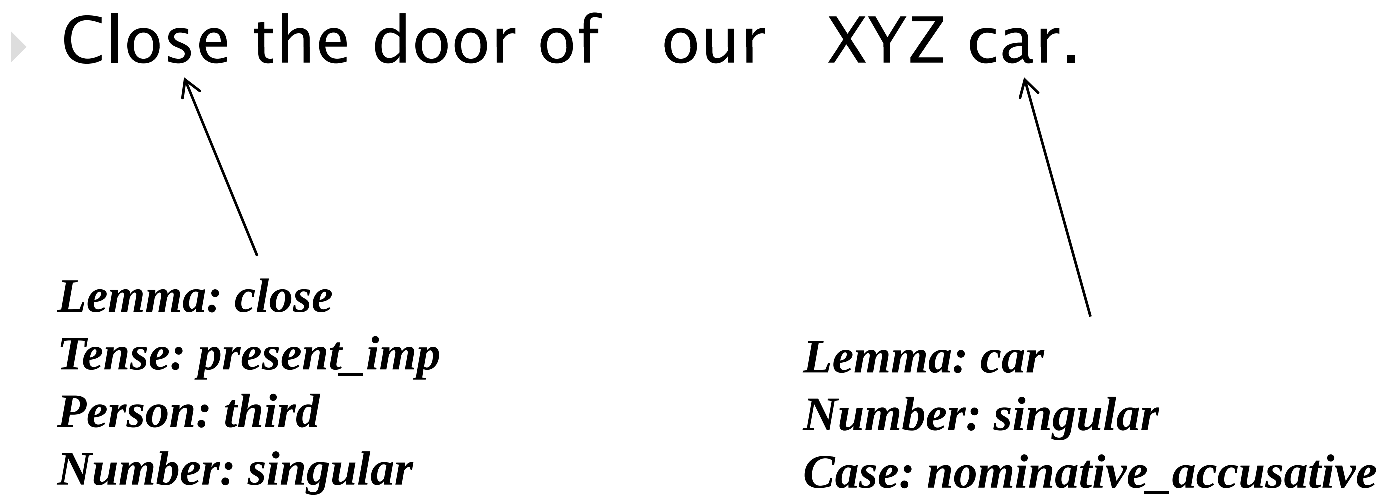
POS tagging, such as:

- Close the door of our XYZ car.
- V DET N PREP PRON NE N

XML and attribut
value structures

statistical methods
large dictionaries

morphology



- Close the door of our XYZ car.

Lemma: close
Tense: present_imp
Person: third
Number: singular

Lemma: car
Number: singular
Case: nominative_accusative

based on dictionaries,
rules for inflection
and derivation

terminology: Why work on terminology?

- Consistency!
- ideally: 1 term = 1 meaning = 1 translation

- less ambiguity, better comprehension, translatability, etc.
- multilingual consistency
- corporate wording

- lower costs (translation but also support)

the reality …

- When analyzing terminology in documents, we find many variants that are used at the same time:
  - web server – web-server
  - upload protection – upload-protection
  - timeout – time out
  - Reset – ReSet
  - sub station – sub-station

how to get consistent terminology?

- author/company defines term banks

- list of deprecated terms
  deprecated term: vehicle
  approved term: car

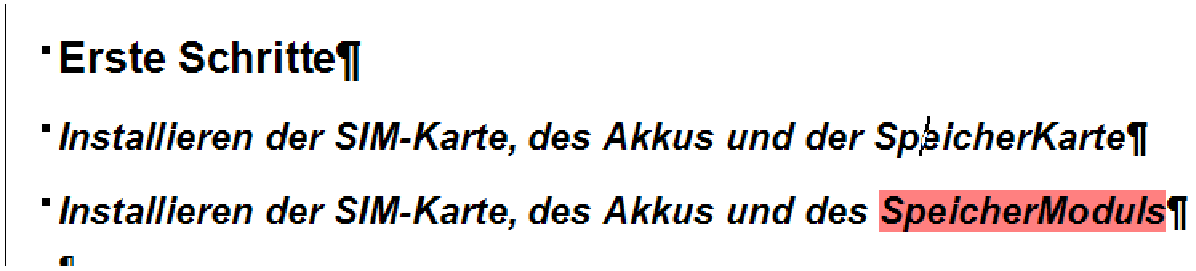
- list of approved terms
  → identification of so-called “variants”
  approved term: SWASSNet User
  deprecated term: SWASSNet user, SWASS-Net User

term variants

- **orthographic variants**
  - hyphen, blank, case: term bank, termbank
- **semi-orthographic variants**
  - number : 6-digit, six-digit
  - trademark : acrolinx IQ™, acrolinx IQ
- **syntactic variants**
  - preposition: oil level, level of oil
  - gerund/noun : call center, calling center
- **synonyms**
  “classical” : vehicle, car
- **language-specific variants**
  (e.g. Fugenelemente DE, Katakana JA)

terminology and spelling

- in terminology: SpeicherKarte



Erste Schritte¶

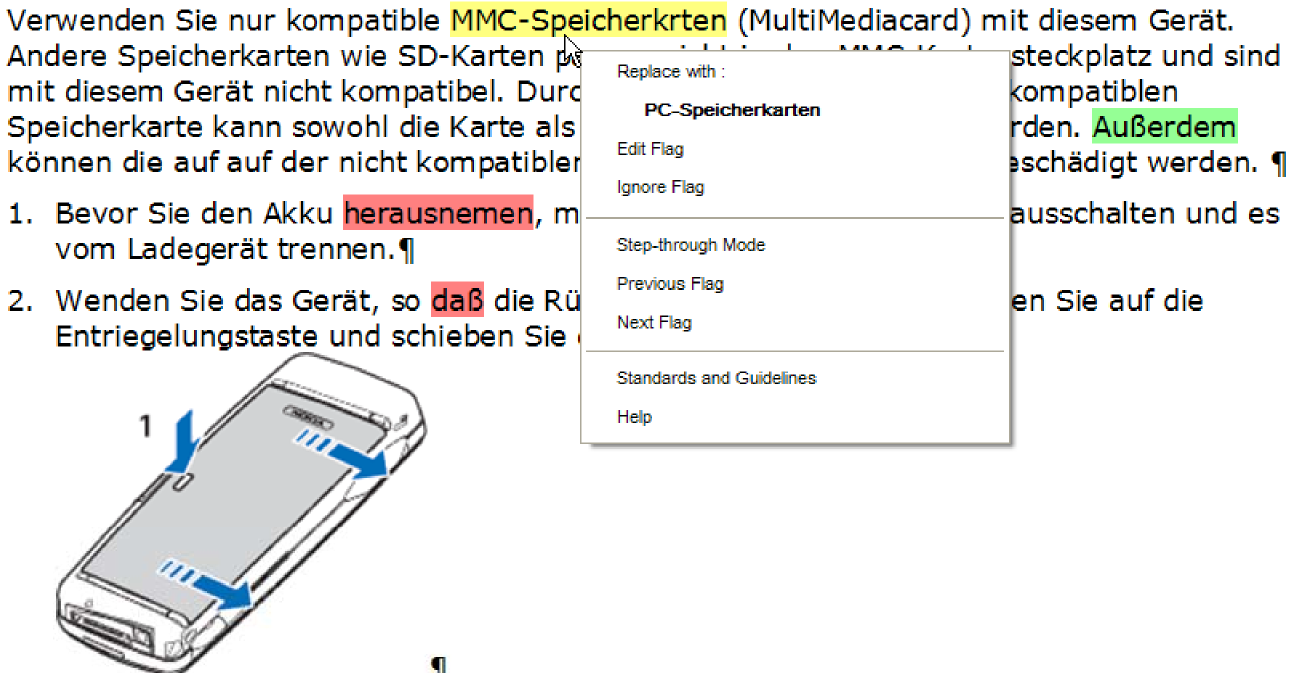
Installieren der SIM-Karte, des Akkus und der SpeicherKarte¶

Installieren der SIM-Karte, des Akkus und des SpeicherModuls¶

¶

terminology and spelling

- term: MMC-Speicherkarten (deprecated), suggested: PC-Speicherkarten

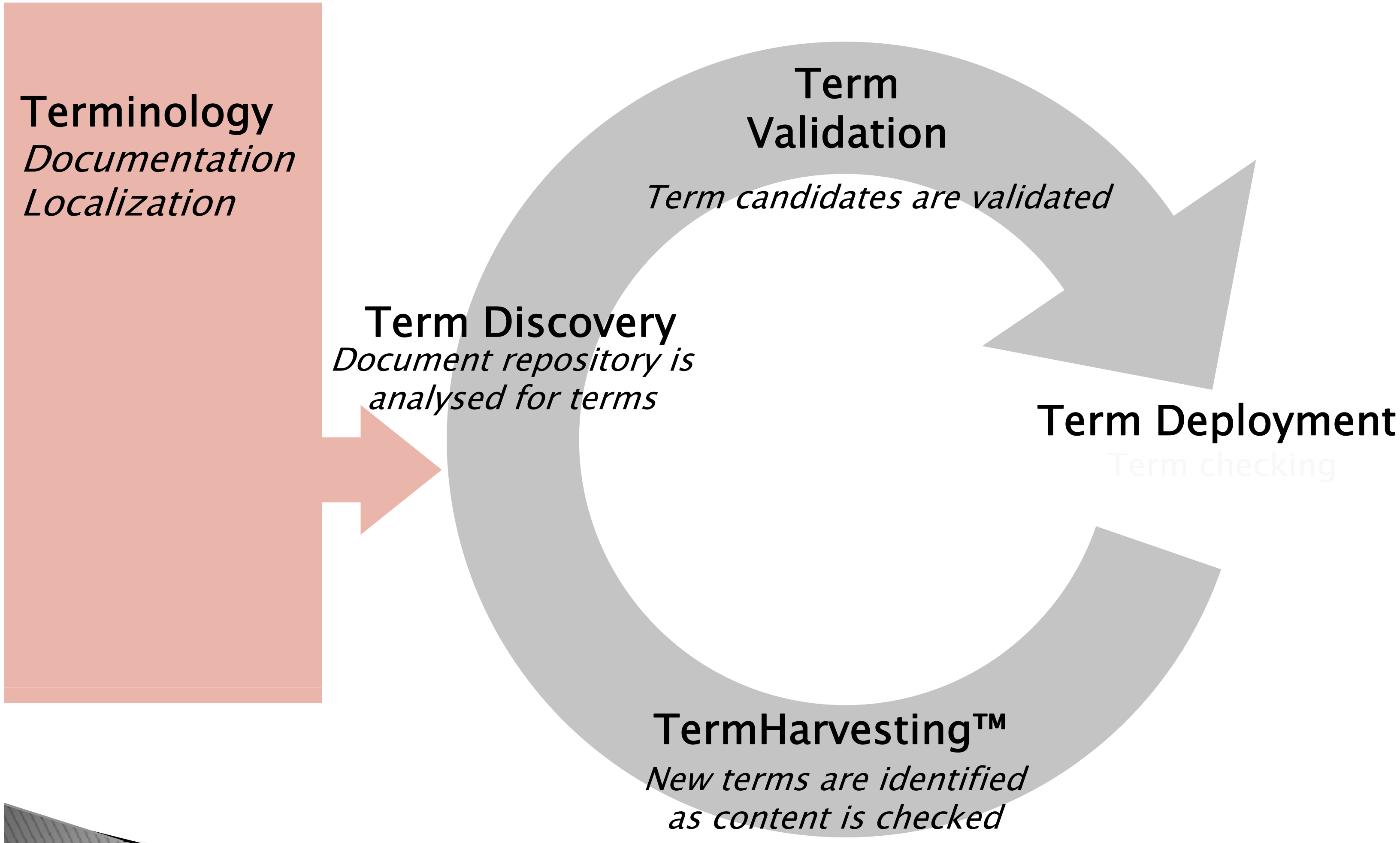


Verwenden Sie nur kompatible MMC-Speicherkrten (MultiMediacard) mit diesem Gerät.
Andere Speicherkarten wie SD-Karten p steckplatz und sind
mit diesem Gerät nicht kompatibel. Durc kompatiblen
Speicherkarte kann sowohl die Karte als rden. Außerdem
können die auf auf der nicht kompatiblen eschädigt werden. ¶

1. Bevor Sie den Akku herausnemen, m ausschalten und es vom Ladegerät trennen.¶
2. Wenden Sie das Gerät, so daß die Rü en Sie auf die Entriegelungstaste und schieben Sie

¶

terminology lifecycle management



Terminology
Documentation
Localization

Term
Validation

Term candidates are validated

Term Discovery
Document repository is analysed for terms

Term Deployment
Term checking

TermHarvesting™
New terms are identified as content is checked

acrolinx
TECHNOLOGY FOR INFORMATION QUALITY

NLP for terminology

- NLP methods for term extraction
  - corpus analysis (morphology, POS, NER)
  - information extraction (potential product names)
  - ontologies (e.g. semantic groups)
- NLP methods for setting up a term database
  - morphology (finding the lemma)
  - POS
- NLP methods for term checking
  - variants
  - similar words
  - inflection

approaches to grammar checking

- definition of correct grammar
  - e.g. HPSG, LFG, chunk-grammar, statistical grammars
  - anything that's not analyzable must be a grammar error
  - preconditions:
    - grammar with large coverage
    - giant dictionaries
    - robust, but not too robust parsing
    - efficient parsing methods
  - high recall, low precision

descriptive grammar

- grammar errors are implemented
  - preconditions:
    - work with error corpora
    - error grammar with a high number of error types
    - „deepness“ of analysis varies with the type of error to be described
  - high precision, recall is based on the number of rules

error grammar

grammar rules, for example

- **subject verb agreement:**
  - Check if instructions are programmed in such a way that a scan never finish.
  - When the operations is completed, the return to home completes.
- **a an distinction:**
  - a isolating transformer
  - an program
- **wrong verb form:**
  - it cannot communicates with them
  - IP can be automatically get

acrolinx
TECHNOLOGY FOR INFORMATION QUALITY

example grammar rule

- **write_words_together**

```
@can ::= [ TOK "^(can)$"
           MORPH.READING.MCAT "^Verb$" ];
```

- The application can not start.
- The application can tomorrow not start.

```
TRIGGER(80) == @can^1 [@adv]* 'not'^2
    -> ($can, $not)
    -> { mark: $can, $not;
         suggest: $can -> ", $not -> 'cannot';
       }
```

- Branch circuits can not only minimize system damage but can interrupt the flow of fault current

```
NEG_EV(40) == $can 'not' 'only' @verbInf   []* 'but';
```

style – controlled language

- controlled languages
  - AECMA – now:
    AeroSpace and Defence Industries Association of Europe (ASD)
    ASD-STE100 (simplified English)
  - Caterpillar Technical English (CTE)
- disadvantage:
  - very restrictive! Prescriptive rules define allowed structures and allowed vocabulary → all other structures and words as disallowed
  - low acceptance of user

style – error definition

- rules define errors (just as grammar rules do)
- rules are defined by user / author
- acceptance is much higher

style

- style guidelines can be different for different usages
  - text type (e.g., press release – technical documentation)
  - domain (e.g., software – machines)
  - readers (e.g., end users – service personnel)
  - authors (e.g., Germans tend to write long sentences)

style rule examples: best practise

- avoid_latin_expressions
- avoid_modal_verbs
- avoid_passive
- avoid_split_infinitives
- avoid_subjunctive
- use_serial_comma
- use_comma_after_introductory_phrase
- spell_out_numerals

style rule examples: company

- use_units_consistently
- abbreviate_currency
- COMPANY_trademark
- do_not_refer_to_COMPANY_intranet
- add_tag_to_UI_string
- avoid_trademark_as_noun
- avoid_articles_in_title

style rule examples MT (pre-editing)

- avoid_nested_sentences
- avoid_ing_words
- keep_two_verb_parts_together
- avoid_parenthetical_expressions

- dependent of MT system and language pair

style rule suggestions

- replacement of words or phrases
- replacement using the correct writing with uppercase or lowercase
- replacement of words using the correct inflection
- generation of whole sentences (e.g. passive – active) requires semantic analysis and generation and is therefore not (yet) possible

example of style rule

- **avoid_future**

- /* Example: „.. It will be necessary .." */

```
TRIGGER (80) == @will^1 [-@comma]* @verbInf^2
                  ->($will, $verbInf)
                  -> { mark : $will, $verbInf;}
```

- /* Example: „.. The router services will be offered in the future .." */

- NEG_EV(40) == $will []* @next @time;

consistent phrasing: why?

- Use the same phrase for the same meaning.

- Examples:
  - Congratulations on acquiring your new wearable digital audio player
  - Congratulations, you have acquired your new wearable digital audio player!
  - Dear Customer, congratulations on purchasing the new wearable digital audio player!

- Using the same phrase makes the documents more readable and helps to save translation costs.

the reality...

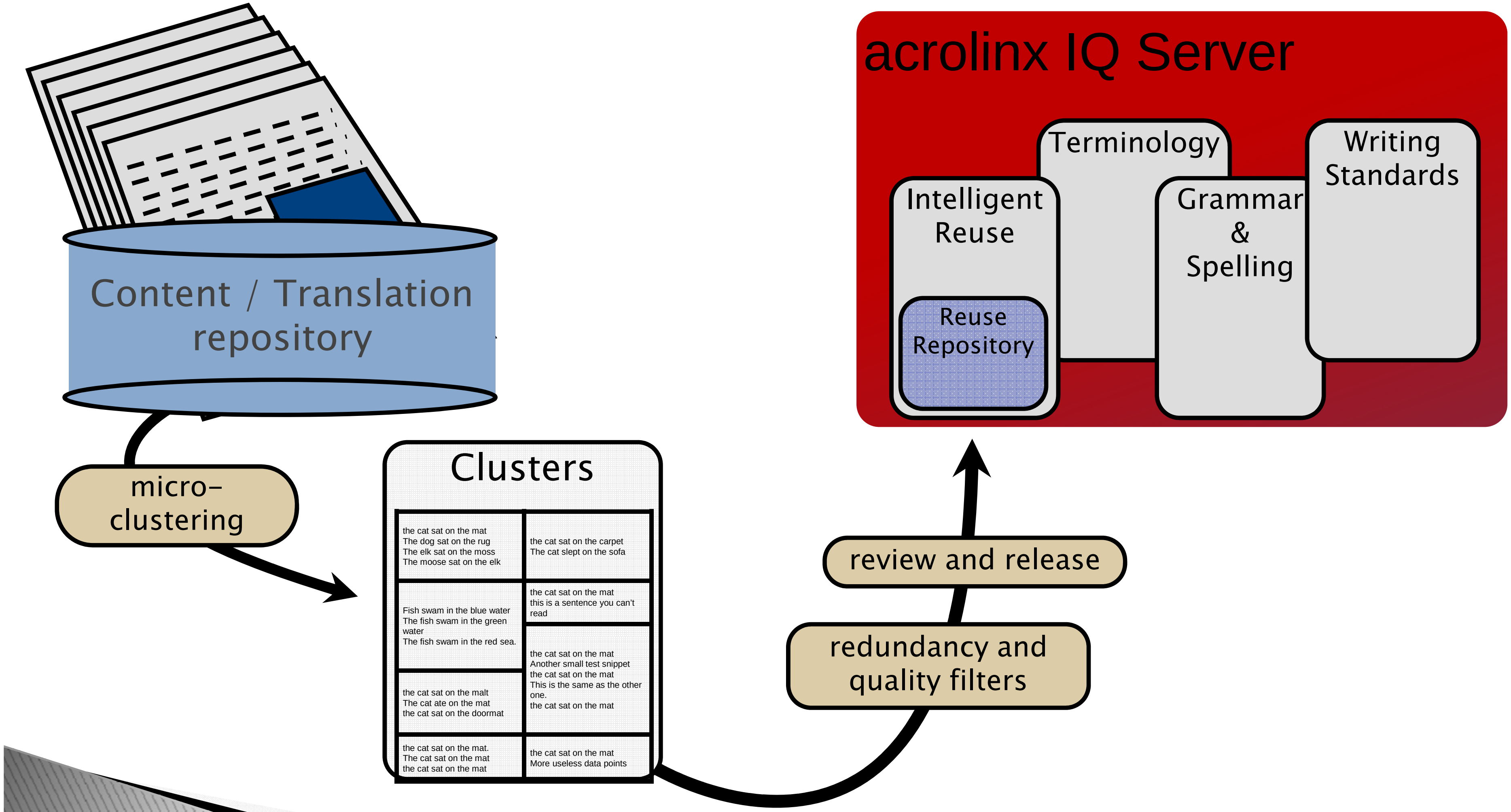
- End date must be equal to or later than the start date.
- End Date must be greater than or equal to Start Date.
- End Date must be greater than Start Date.
- End Date must be later than Start Date.
- End date should be greater than start date.
- End Date cannot be before the Start Date.
- Please enter an end date that is later than the start date.
- Please enter an End Date that is later than or the same as the Start Date.
- Please enter a start date that is before the end date.
- Start date must be before end date!
- The end date must be later than or the same as the start date.
- The start date cannot be later than the end date.
- The start date must be on or before the end date.
- The Start Date cannot be after the End Date.
- The end date cannot be before the start date.
- The actual end date must be on or after the actual start date.
- The start date must be prior to the end date.
- The ending date must be later than or the same as the beginning date.
- Your end date must be after your start date.
- You cannot enter an "End Date" that is before your "Start Date."
- Your start date must be before your end date.
- You entered a start date later than the end date.

Intelligent Reuse

- analysis of text documents with NLP, such as ontologies, morphology, sentence similarity
- selection of standard sentences
  - automatic selection with respect to grammar, style, terminology
  - human validation
- suggestions for similar sentences in new texts

Intelligent Reuse™

Building your Reuse Repository



the cat sat on the mat
The dog sat on the rug
The elk sat on the moss
The moose sat on the elk

Fish swam in the blue water
The fish swam in the green water
The fish swam in the red sea.

the cat sat on the malt
The cat ate on the mat
the cat sat on the doormat

the cat sat on the mat.
The cat sat on the mat
the cat sat on the mat

the cat sat on the carpet
The cat slept on the sofa

the cat sat on the mat
this is a sentence you can’t read

the cat sat on the mat
Another small test snippet
the cat sat on the mat
This is the same as the other one.
the cat sat on the mat

the cat sat on the mat
More useless data points

acrolinx
TECHNOLOGY FOR INFORMATION QUALITY

NLP components in acrolinx

- components for analysis
  - tokenizer (sentences and words)
  - morphology, decomposition
  - POS tagger
  - word guesser
  - gazetteer

NLP components in acrolinx

- rule formalism is based on language analysis results
  - spelling
  - grammar
  - style
  - term variants
  - term extraction

Find out more at our Knowledge Center

www.acrolinx.com

acrolinx
TECHNOLOGY FOR INFORMATION QUALITY