gold standards and evaluation

- a gold standard is a training and evaluation resource
- gold standards contain expert (-checked) annotations
- typical machine learning evaluation scenario: n-fold cross-validation
  - divide the resource into n equal parts
  - train on n - 1 of the parts, test on the remaining part
  - do this n times, so that testing is done using each of the n parts
  - take the average of the n results (precision, recall, F-score, accuracy, etc.) as the result of the evaluation

gold standard issues (again)

- gold standards are often post hoc, originally developed for other purposes than evaluation (e.g., the Penn treebank; WordNet)
- gold standards are convenience samples wrt genre, modality, annotation format and content
- there has been very little systematic investigation of the influence of, e.g., genre, text type, language, etc., on the accuracy
- who are the experts? – crowdsourcing approaches increasingly common (e.g., Amazon’s Mechanical Turk)

LT tools – the CLARIN taxonomy, 1

- annotation tool
- evaluation tool
- multimedia tool
- multimodal tool
- spoken ∼ written
- NLP development aid
- single ∼ multiple task
- toolbox
- web service
- web application
LT tools – the CLARIN taxonomy, 2

- user categories:
  - language technology developers
  - software integrators
  - pipe creators
  - informed end users
  - naive end users

LT tools – the CLARIN taxonomy, 3

- text:
  - analysis
  - information retrieval
  - information extraction
  - authoring aids/ knowledge representation
  - data storage
  - translation
  - modality conversion

- speech:
  - analysis
  - text-to-speech
  - spoken language translation
  - speech-to-text
  - spoken language understanding
  - spoken dialogue systems

LT tools – the CLARIN taxonomy, 4

- lexical resources:
  - lookup/ browse
  - lexical acquisition
  - maintenance/ authoring
  - analysis
  - lexicon management

- corpora:
  - editing
  - indexing
  - annotation
  - alignment
  - concordancer
  - querying

LT tools – the CLARIN taxonomy, 5

(From CLARIN deliverable D5C-2, p. 45)
LT processing chains

- LT processing chains are ad-hoc combinations of LT tools...
- (which may draw on LT resources)
- ...for one-off tasks
- or for building prototypes
- (cf. Unix command-line chains:
  - `cat text.txt | tr '[[:punct:]]' ' ' | tr -s ' ' '\n' | sort | uniq -c | sort -nr -k 1,1 > wfrq.txt
- )
- LT processing chains require both data standardization and content model standardization (to be discussed later)

LT processing chains and web services

- web services are programs run using the internet as communication channel
- two standard protocols: SOAP (Simple Object Access Protocol; XML) and REST (REpresentational State Transfer; http commands)
- WSDL (Web Services Description Language; XML) communicates capabilities and requirements of web services
- web service based LT processing chains simply implement each LT module as a web service

A web service based LT processing chain

WebLicht (from CLARIN deliverable D5R-3a, p. 8)

web services – not only chaining

- note that web services are general processing modules
- similar to functions/ subroutines in programming languages
- hence, they could in principle support other processing architectures than pipelines, e.g.
  - blackboard architectures
  - dataflow architectures
  - event-driven architectures
BLARK stands for Basic Language Resource Kit: the minimal set of language resources that is necessary to do any precompetitive research and education at all (Krauwer 2003).

- The acronym was coined in connection with a concerted resource-building effort for Dutch.
- A BLARK is defined with a set of task domains.
- Existing tools and resources are surveyed.
- Tools and resources are ranked with the number of domains they serve (the BLARK matrix).

### BLARK Task Domains

- Computer-assisted language learning
- Access control
- Speech input
- Speech output
- Dialogue systems
- Document production
- Information access
- Translation

### BLARK Elements: Text

- Robust text pre-processing
- Morphological analysis
- Syntactic analysis
- Semantic analysis
- Monolingual lexicon
- Bilingual lexicons
- Written treebank
- Parallel corpora (annotated, aligned)
- Evaluation benchmarks

### BLARK Elements: Speech

- Automatic speech recognition
- Speech synthesis
- Calculation of confidence measures
- Identification tools
- Tools for (semi-)automatic annotation of speech corpora
- Multi-modal speech corpora
- Multi-medial speech corpora
- Multi-lingual speech corpora
- Evaluation benchmarks
The BLARK matrix

<table>
<thead>
<tr>
<th>Methods</th>
<th>Data</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>訓語素化</td>
<td></td>
</tr>
<tr>
<td></td>
<td>巻頭選句</td>
<td></td>
</tr>
<tr>
<td></td>
<td>名詞位置付け</td>
<td></td>
</tr>
<tr>
<td></td>
<td>文節分割</td>
<td></td>
</tr>
<tr>
<td></td>
<td>名詞語素分析</td>
<td></td>
</tr>
<tr>
<td></td>
<td>詞性分け</td>
<td></td>
</tr>
<tr>
<td></td>
<td>文法変換</td>
<td></td>
</tr>
<tr>
<td></td>
<td>文法変換解析</td>
<td></td>
</tr>
<tr>
<td></td>
<td>語彙解析</td>
<td></td>
</tr>
<tr>
<td></td>
<td>文法変換生成</td>
<td></td>
</tr>
<tr>
<td></td>
<td>言語学的変革</td>
<td></td>
</tr>
<tr>
<td></td>
<td>英文版</td>
<td></td>
</tr>
<tr>
<td></td>
<td>英語版</td>
<td></td>
</tr>
<tr>
<td></td>
<td>仏語版</td>
<td></td>
</tr>
<tr>
<td></td>
<td>仏語版</td>
<td></td>
</tr>
</tbody>
</table>

BLARKs can be everywhere

- the BLARK notion and methodology is widely applicable
- generalized, it consists of the elements:
  1. define task domains
  2. survey resources
  3. determine dependence of task domains upon resources
  4. thus achieving a ranking of resources
  5. build/acquire missing resources
- the translation and (multilingual) information access task domains are relevant, e.g., in working with historical language stages (see the diachronic BLARK paper)