Assessing the Quality of Natural Language Text

DC Research Ulm (RIC/AM)

daniel.sonntag@dfki.de

GI 2004

Agenda

- Introduction and Background to Text Quality
- Text Quality Dimensions
  - Intrinsic Text Quality, Accessibility, Contextual and Representational Text Quality
- Text Quality Metrics
  - Text Quality Audit Guideline
  - Computational Metrics
- Application example

Introduction and Background

- Empirical approach: Data quality -> Text quality
- Data Quality:
  - Structured data fields: numbers, names, customer address, ...
  - Correct data entries, no duplicates, referential integrity (RDBS)
  - Correct process models (data warehouse), **Data Cleaning**
- Text Quality:
  - unstructured data fields: Free text of arbitrary length
  - Examples: reports, customer emails, web pages, papers, ...

Introduction and Background

- Linguistic data are ubiquitous.
- Data Retrieval -> Information Retrieval
- Information Retrieval Market -> NLP Market
- Data Quality -> Text Quality

Text Quality Dimensions

- Concept: *Fitness for use, user judges whether data is fit for use, beyond accuracy.*
- Basis: Data Quality Dimensions [WS96]

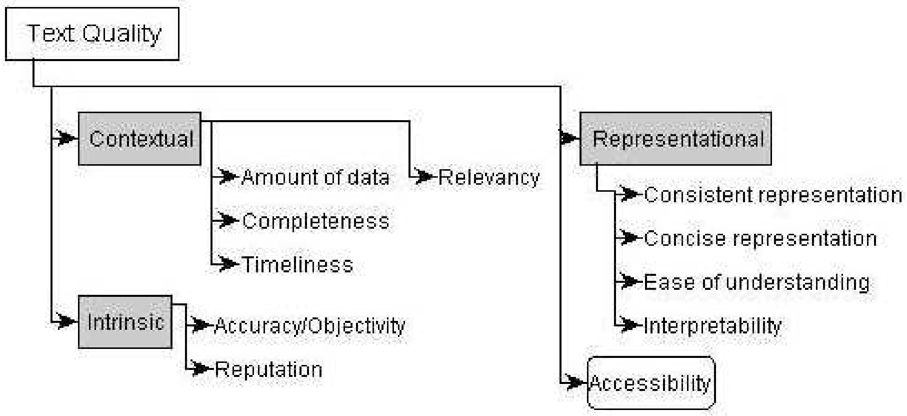
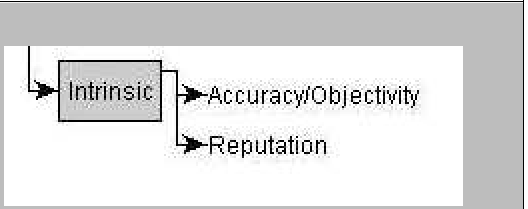


Figure 1: Text quality dimensions.

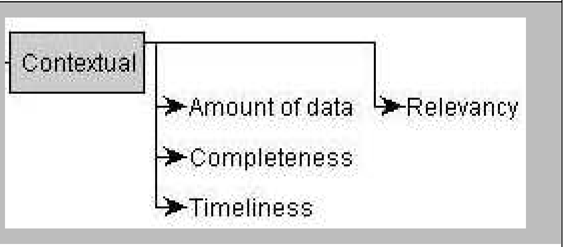
Text Quality Dimensions



- Intrinsic Text Quality
  - M: No comp. to other documents content is possible.
  - H/M: Reputation very common for textual domain.
  - Example: author, institution, references, print run in research papers
    - transitive to references contained

- reputation ~~<---->~~ believability ~~<---->~~ accuracy

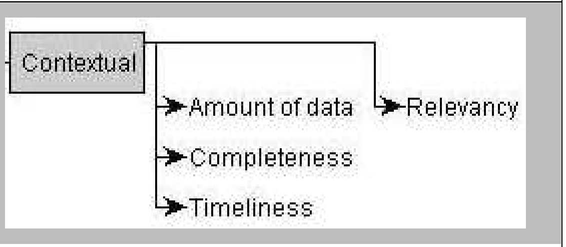
Text Quality Dimensions



- Contextual Text Quality
  - Common Approach: Appropriate contextual parameters must be set manually.
  - Example:
    - Weather report: short, concise, no complete sentences
    - Pilot manual: restricted vocabulary
- Suitability for automatic processing

- **Contextual -> Representational Text Quality**

Text Quality Dimensions



Contextual

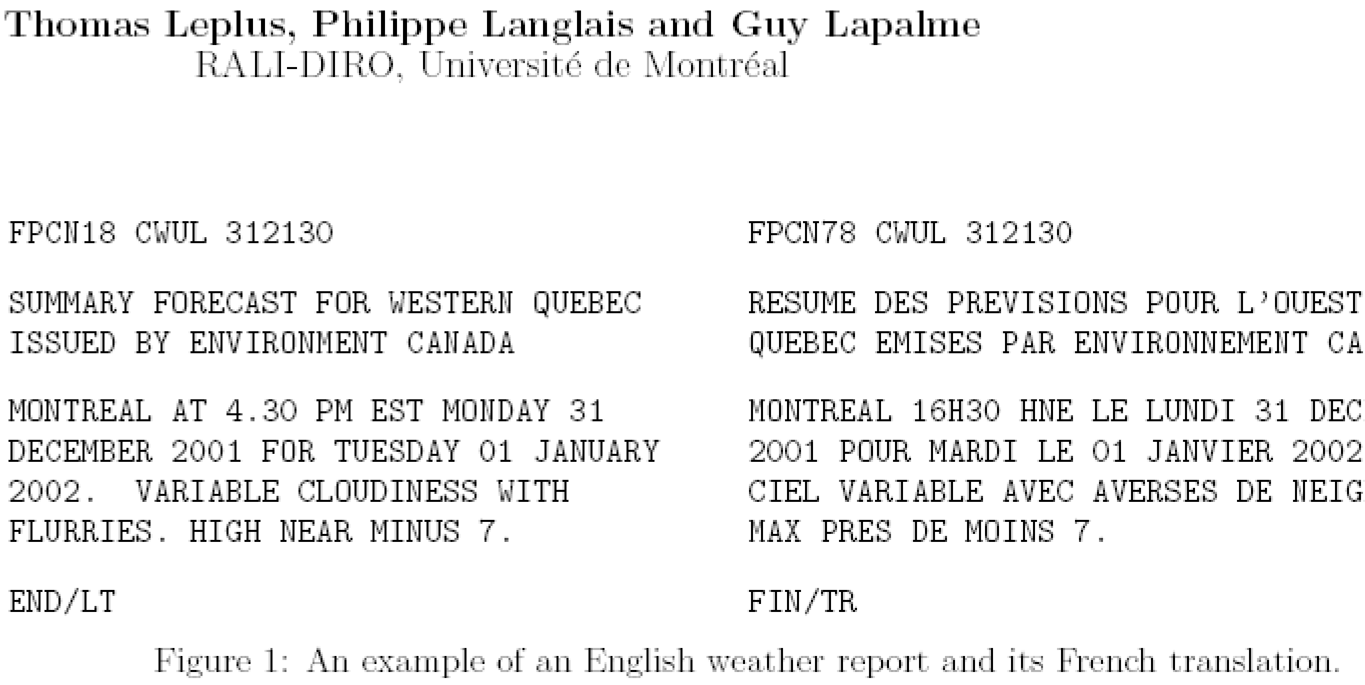
Amount of data

Relevancy

Completeness

Timeliness

- Weather Report Example

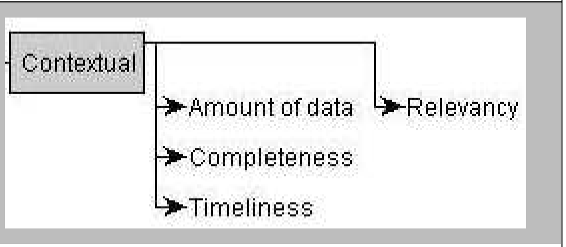


Thomas Leplus, Philippe Langlais and Guy Lapalme
RALI-DIRO, Université de Montréal

```
FPCN18 CWUL 312130                      FPCN78 CWUL 312130

SUMMARY FORECAST FOR WESTERN QUEBEC     RESUME DES PREVISIONS POUR L'OUEST DU
ISSUED BY ENVIRONMENT CANADA            QUEBEC EMISES PAR ENVIRONNEMENT CANADA

MONTREAL AT 4.30 PM EST MONDAY 31       MONTREAL 16H30 HNE LE LUNDI 31 DECEMBRE
DECEMBER 2001 FOR TUESDAY 01 JANUARY    2001 POUR MARDI LE 01 JANVIER 2002.
2002.  VARIABLE CLOUDINESS WITH         CIEL VARIABLE AVEC AVERSES DE NEIGE.
FLURRIES. HIGH NEAR MINUS 7.            MAX PRES DE MOINS 7.

END/LT                                  FIN/TR
```

Figure 1: An example of an English weather report and its French translation.

Text Quality Dimensions



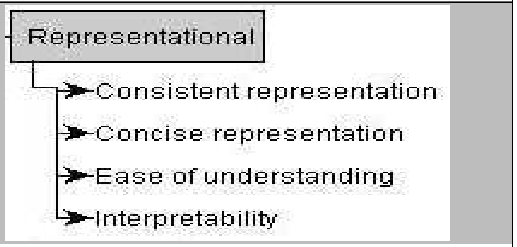
- Pilot Manual Example
  - Technical Writing Guideline:
    - *Be as specific and simple as possible. Use simple sentence construction.Identify nouns with specific names. Take advantage of labels on equipment diagrams. Define terms which may be unfamiliar to your reader, or leave them out. Avoid using "a lot", "very", "much", and "significantly“, ...*
    - http://research.faa.gov/aar/tech/docs/techreport/01-43.pdf

–New procedures are reviewed by a peer, approved by a lead writer, and proofread for grammar and typographic accuracy.

– Vocabulary is limited to include only words found in the United States Air Force dictionary.

– The writing process is ISO 9000 certified.

Text Quality Dimensions

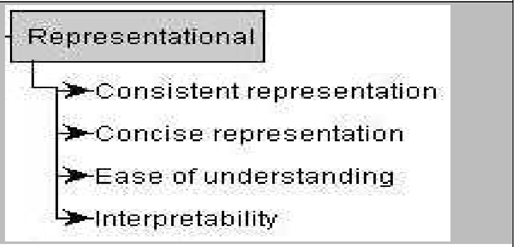


Representational

- Consistent representation
- Concise representation
- Ease of understanding
- Interpretability

- Representational Text Quality
- Explain by Representational deficits:
  - **Single Source Problems**:
    - Wrong formulated data values, typing errors, different spellings of same word, co-reference problems, lexical ambiguity.
    - Wrong formulated data values: 3. October, 3rd of October, October 3rd.
    - Co-reference problem: *Peter = he = my friend*
    - Lexical ambiguity: *Birne*

Text Quality Dimensions



Representational

- Consistent representation
- Concise representation
- Ease of understanding
- Interpretability

- Representational Text Quality
  - **Multi-Source Problems**:
    - Homonym name conflicts (special type of lexical ambiguity)
    - Document duplicates (and identification)

Text Quality Metrics

- Assessing Text Quality -> Data Cleaning <- Obtain and interpret metric values automatically.
- Problems:
  - Lack of methods to get at the real content of texts vs.

  Content defines very important quality measures.

  -> **Many reservations on effective quality measures.**
- Idea:
  - Define Text Quality with respect to consumer, human and machine separately.
  - Focus on Text Representation for classification, retrieval and IE.

Text Quality Metrics

- Audit guideline: Measure inconformities: Direct observation vs. Reading the text.
- Introduce new viewpoint.
  - Condition context: (1) text to satisfy information need, (2) info contained in text, (3) info candid, (4) info can be extracted by native speaker

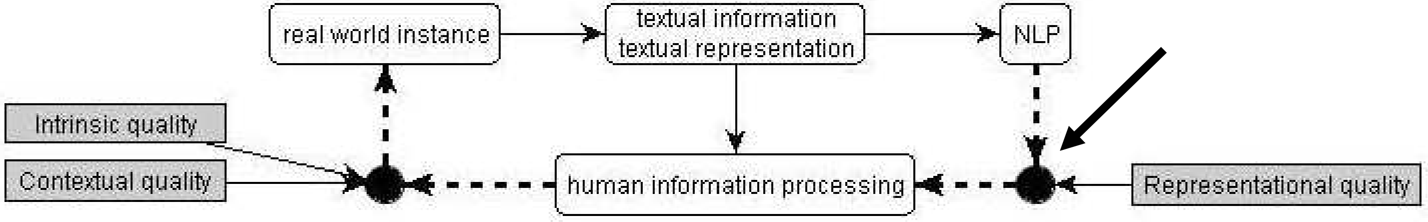
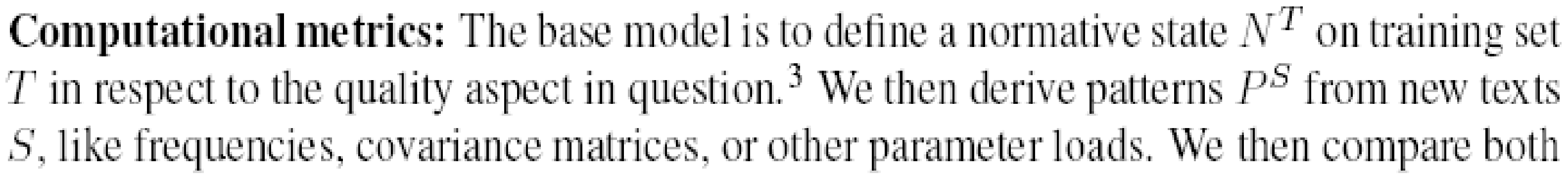


Figure 2: Text quality audit guideline.

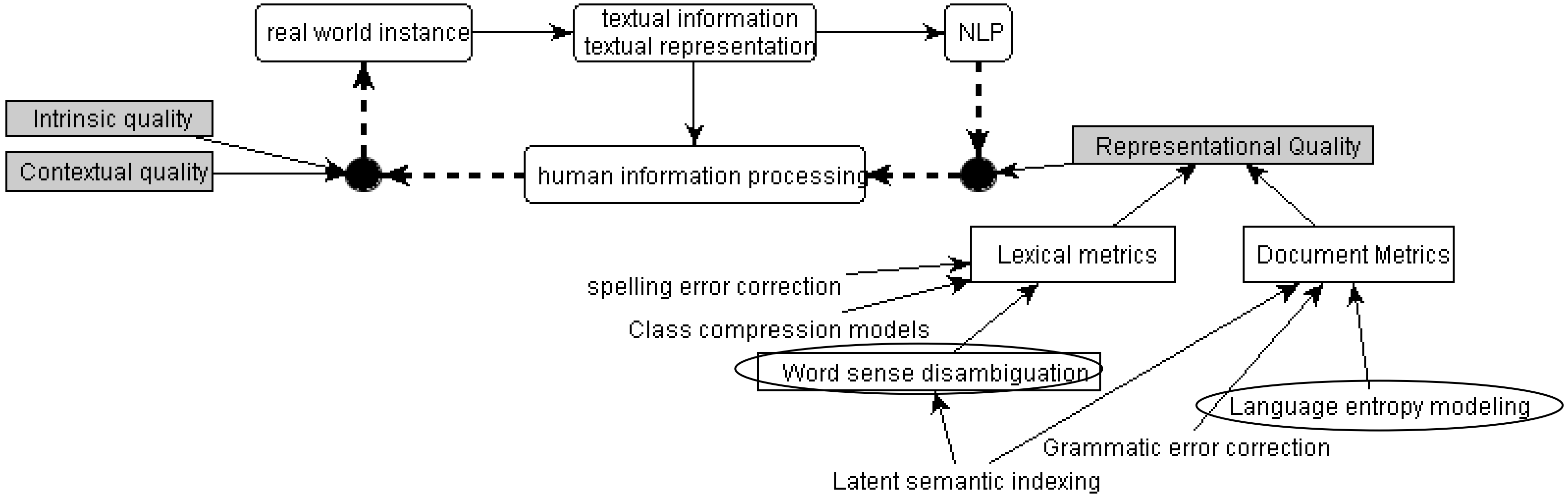
Computational Metrics



Computational metrics: The base model is to define a normative state N^T on training set T in respect to the quality aspect in question.[3] We then derive patterns P^S from new texts S, like frequencies, covariance matrices, or other parameter loads. We then compare both

- degree of deviation $\delta(N^T, P^S)$
- The less the degree, the better the quality.

Computational Metrics



- Spelling quality: frequency patterns (error/appearance)
- Ambiguity quality: Homonym lookup in synonym classes
- Grammatical quality: error frequen*The dogs is barking.*

Computational Metrics

- Language Entropy Modelling

$$H(W_{1,n}) = -\sum_{w_{1,n}} P(w_{1,n}) \log P\,(w_{1,n})$$

  - Entropy: measure of uncertainty about what a message says.
  - Uncertainty can be regarded as lack in quality.

- Figure of merit to compare models: per word **cross entropy**

$$\frac{1}{n} H\,(W_{1,n}) = -\frac{1}{n} \sum_{w_{1,n}} P\,(w_{1,n}) \log \underline{\;P_{Model}\;}\,(w_{1,n})$$

$$H\,(L, P_{Model}) = -\lim_{n \to \infty} \frac{1}{n} \sum_{w_{1,n}} \log\ P_{Model}\,(w_{1,n})$$

Computational Metrics

- Language Entropy Modelling
  - The model with smallest cross entropy is best of the lot.

degree of deviation $\delta(N^T, P^S)$

$$H\,(W_{1,n}) \leq H\,(W_{1,n} P_{Model}\ )$$

  - (Language must be *ergodic*.)

Computational Metrics + Enhancement

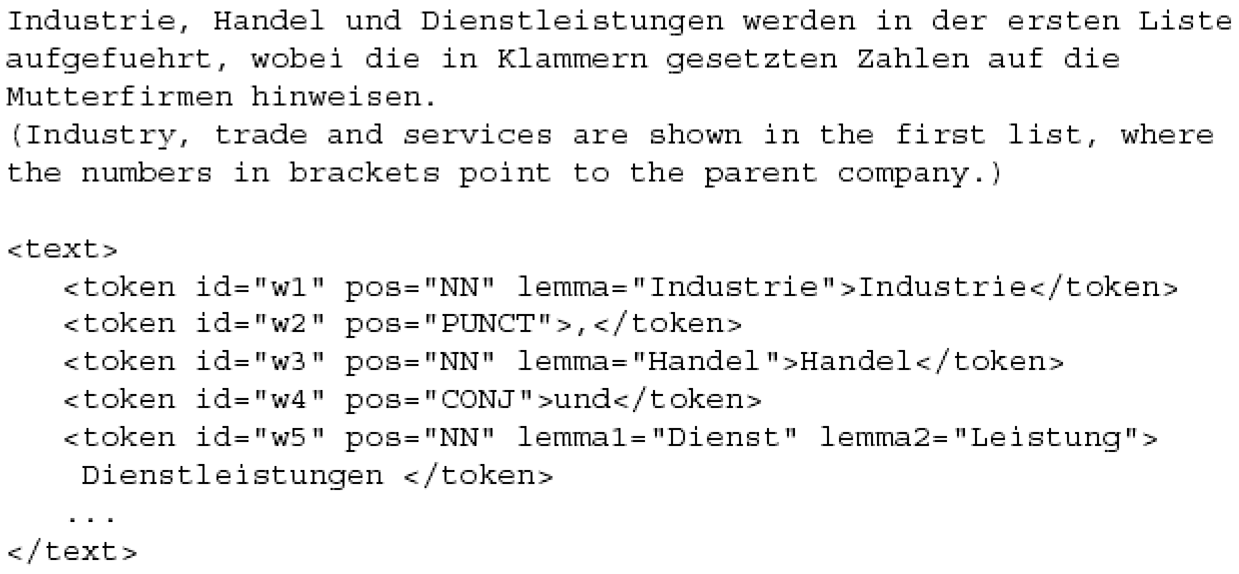
- Word Sense Disambiguation + Annotating word sense
  - Automatically: Latent Semantic Indexing
  - Manually: Rely on annotation, rely on better **content** representation
    - -> Semantic Web Activity

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation." -- *Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001*

- - -> Linguistic Annotation for the Semantic Web
  - -> SmartWeb

SmartWeb is based on two parallel efforts that have the potential of forming the basis for the next generation of the Web. The first effort is the semantic Web which provides the tools for the explicit markup of the content of Web pages; the second effort is the development of semantic Web services

Application: Linguistic Annotation for the Semantic Web

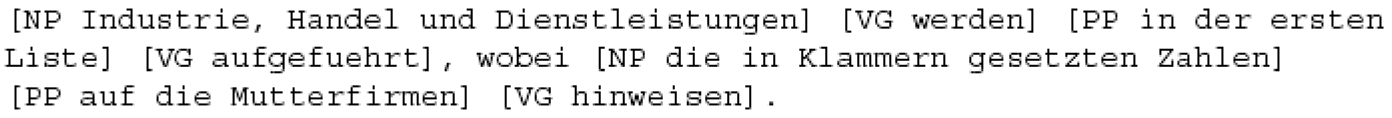


Industrie, Handel und Dienstleistungen werden in der ersten Liste aufgefuehrt, wobei die in Klammern gesetzten Zahlen auf die Mutterfirmen hinweisen.
(Industry, trade and services are shown in the first list, where the numbers in brackets point to the parent company.)

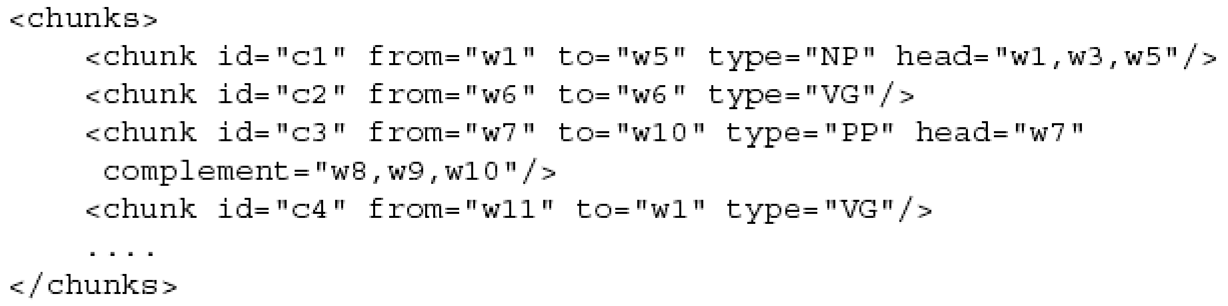
```
<text>
   <token id="w1" pos="NN" lemma="Industrie">Industrie</token>
   <token id="w2" pos="PUNCT">,</token>
   <token id="w3" pos="NN" lemma="Handel">Handel</token>
   <token id="w4" pos="CONJ">und</token>
   <token id="w5" pos="NN" lemma1="Dienst" lemma2="Leistung">
    Dienstleistungen </token>
   ...
</text>
```

Example from: Paul Buitelaar and Thierry Declerck DFKI GmbH

Application: Linguistic Annotation for the Semantic Web



```
[NP Industrie, Handel und Dienstleistungen] [VG werden] [PP in der ersten
Liste] [VG aufgefuehrt], wobei [NP die in Klammern gesetzten Zahlen]
[PP auf die Mutterfirmen] [VG hinweisen].
```



```
<chunks>
    <chunk id="c1" from="w1" to="w5" type="NP" head="w1,w3,w5"/>
    <chunk id="c2" from="w6" to="w6" type="VG"/>
    <chunk id="c3" from="w7" to="w10" type="PP" head="w7"
     complement="w8,w9,w10"/>
    <chunk id="c4" from="w11" to="w1" type="VG"/>
    ....
</chunks>
```

Example from: Paul Buitelaar and Thierry Declerck DFKI GmbH

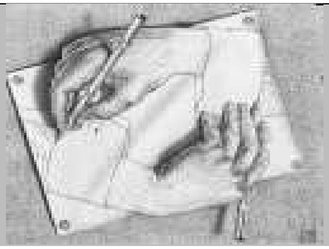
Application: SmartWeb

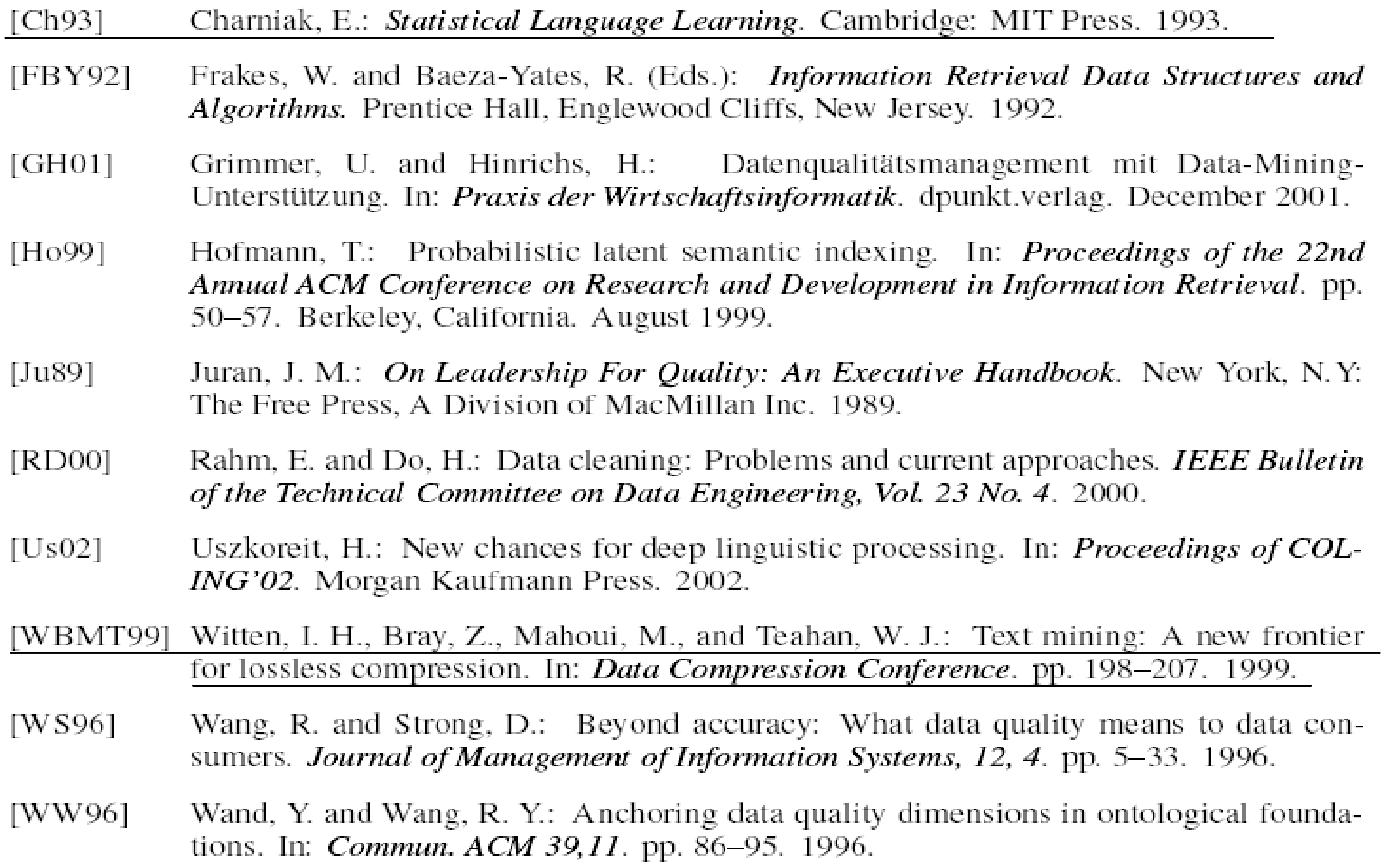
- SmartWeb: Mobile Applications of the Semantic Web
- Two parallel efforts to form basis for the next generation of the Web:
  - Provide tools for explicit markup of content of Web pages.
  - Development of *Semantic Web Services*, agents become producers and consumers of information (^= **automatic NLP**).

Conclusions

- Data Quality dimensions suitable for Text Quality dimensions.
- Measure Text Quality for Humans and NLP separately.
- Representational Text Quality is measurable.
- Text quality for NLP traces back to questions of text representation.
- **Better text representation is the key for better text quality.**

Reading Material





[Ch93] Charniak, E.: *Statistical Language Learning*. Cambridge: MIT Press. 1993.

[FBY92] Frakes, W. and Baeza-Yates, R. (Eds.): *Information Retrieval Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, New Jersey. 1992.

[GH01] Grimmer, U. and Hinrichs, H.: Datenqualitätsmanagement mit Data-Mining-Unterstützung. In: *Praxis der Wirtschaftsinformatik*. dpunkt.verlag. December 2001.

[Ho99] Hofmann, T.: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*. pp. 50–57. Berkeley, California. August 1999.

[Ju89] Juran, J. M.: *On Leadership For Quality: An Executive Handbook*. New York, N.Y: The Free Press, A Division of MacMillan Inc. 1989.

[RD00] Rahm, E. and Do, H.: Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 23 No. 4*. 2000.

[Us02] Uszkoreit, H.: New chances for deep linguistic processing. In: *Proceedings of COLING'02*. Morgan Kaufmann Press. 2002.

[WBMT99] Witten, I. H., Bray, Z., Mahoui, M., and Teahan, W. J.: Text mining: A new frontier for lossless compression. In: *Data Compression Conference*. pp. 198–207. 1999.

[WS96] Wang, R. and Strong, D.: Beyond accuracy: What data quality means to data consumers. *Journal of Management of Information Systems, 12, 4*. pp. 5–33. 1996.

[WW96] Wand, Y. and Wang, R. Y.: Anchoring data quality dimensions in ontological foundations. In: *Commun. ACM 39,11*. pp. 86–95. 1996.

Questions