INTERLINGUAL ASPECTS OF WIKIPEDIA'S QUALITY[1]

(Research-in-progress)

Rainer Hammwöhner
U. of Regensburg, Germany
rainer.hammwoehner@sprachlit.uni-regensburg.de

Abstract: This paper presents interim results of an ongoing project on quality issues concerning Wikipedia. One focus of research is the relation of language and quality measurement. The other one is the use of interlingual relations for quality assessment and improvement. The study is based on mono- and multilingual samples of featured and non-featured Wikipedia articles in English, French, German, and Italian that are evaluated automatically.

Key Words: Information Quality, Wikipedia, Knowledge Organization, Indexing

1. INTRODUCTION

The relevance of Wikipedia as an information source for an ever increasing number of users needs not to be stressed any more. This information seeking behavior is backed by a couple of studies concerning the quality of Wikipedia articles. These studies however cover only a limited number of quality aspects that are relevant for the different kinds of conceivable Wikipedia usage.

This paper presents results of a long term project devoted to research on the quality of Wikipedia. The effort invested into this project is motivated by the crucial role of Wikipedia as pilot application for the further development of the Web 2.0. Within the scope of the project a comprehensive model of Wikipedia quality will be developed and a number of empirical studies focusing on selected quality dimensions will be conducted. In the following the research context of the project will be presented first. Then the general idea of the project will be discussed so that a more specific research problem can be derived from this very idea. The design of a couple of small studies will be developed, results will be presented.

1.1 *Important features of Wikipedia*

Wikipedia is the world's largest online encyclopedia offering information in more than 200 languages [24]. It is built and maintained by the cooperative effort of an open community of users and operated by the Wikimedia Foundation. Its technical core is the MediaWiki software. The English Wikipedia contains more than 2 million articles. 15 Wikipedias provide more than 100,000, further 56 more than 10,000 lemmata.

Since most of the readers of this paper will know Wikipedia from own experience, some few additional remarks on central features of this online encyclopedia should be sufficient:

- Authors of Wikipedia may identify themselves by their name or a pseudonym but are not required to do so.

[1] A slightly modified version of this paper has been presented at ICIQ 2007 conference, it is accessible via http://mitiq.mit.edu/iciq/ICIQ/iqdownload.aspx?ICIQYear=2007&File=INTERLINGUAL+ASPECTS+OF+WIKIPEDIAS+QUALITY.pdf

- Administrators are users with extended competences. They may delete articles, freeze article versions or even ban users. Administrators are elected by the communitiy.
- An efficient versioning system prevents any edits from being lost. Every version of an article can be addressed via hyperlink. Unwanted changes to an article can be dealt with by reverting to a previous version.
- Corresponding articles in different languages – e.g. Deutschland, Germany, Allmagne, Tyskland, etc. – are connected by interlanguage links such that users may navigate easily between them in order to learn about culture specific points of view or simply to check the contents of an article.
- Articles are indexed with descriptors from a structured vocabulary – Wikipedia's so called category system. The terms of the category system are arranged to a directed graph. Cycles may occur within this graph.
- Authors of Wikipedia articles may communicate via specific talk pages associated to every article and to the homepages of authenticated users.
- Wikipedia articles with outstanding quality are elected by the community as featured articles. The implicit quality model behind this choice is made explicit by Stvilia et al. [19].

1.2 *Quality models for Wikipedia*

A couple of studies on the subject of Wikipedia's quality has already been published. These studies however cover only some of the quality dimensions relevant to the assessment of an electronic encyclopedia like Wikipedia. In the following these studies will be grouped according to an explicit quality model. Their strengths and weaknesses will be analyzed with respect to that model.

A quality model for Wikipedia may be derived from a general model of information quality, which is part of the AIMQ methodology [13]. Four categories of quality dimensions may be discerned according to this model:

a. *Intrinsic information quality* can be attributed to information without reference to any kind of context. Accuracy, believability, reputation and objectivity may be seen as quality dimensions belonging to this category.
b. *Contextual information quality* refers to the fact that information must be useful for users with respect to questions asked or goals to be accomplished. Thus, it must be relevant, in time, understandable and complete.
c. *Representational information quality* can be attributed to the structures of knowledge organization employed. These structures must be consistent, easy to understand and to manipulate. They must support the context oriented selection of relevant information.
d. *Accessibility information quality* is the quality of the tools provided. Utility, usability and security aspects may be identified.

There are three more specific models, which should be considered and set into relation to the general model mentioned above.

a. Wikipedia has its own quality model, which is explicated at various locations of the encyclopedia. Most aspects of *intrinsic information quality* are defined in the context of the selection of *featured articles* [25]. Featured articles are expected to be well-written, comprehensive, factual accurate and verifiable, neutral or uncontroversial, stable and compliant to Wikipedia manuals. They should be of reasonable length and should contain an appropriate amount of images. Similarly criteria for featured pictures [26], lists [27] and portals [28] are provided and advice for the appropriate use of categories is given [29]. The latter cover aspects of *representational information quality*. Dimensions of *contextual information quality* are addressed only indirectly – since no explicit reader model exists within Wikipedia. *Accessibility information quality* is not addressed at all. The use of the MediaWiki software and its current features are not taken into consideration within this context.
b. A comprehensive quality model for encyclopedias is introduced by Crawford [1]. Eight general quality dimensions are proposed: scope (focus or purpose, subject coverage, audience,

arrangement and style), format, uniqueness, authority, accuracy (accuracy and reliability, objectivity), currency, indexing, relevance to user needs and costs. Criteria like scope and accuracy aim at intrinsic quality aspects. They are quite similar to the quality dimensions proposed for Wikipedia. Coverage, however, is a criterion that does not aim at the single lemma, but at the encyclopedia as a whole. This kind of holistic view is not present in the Wikipedia quality model. Contextual information quality is covered by the quality dimensions currency, relevance to user needs and costs. Indexing seems to aim at some kind of representational information quality. Usability aspects once more are not addressed.

c. Stvilia, Gasser and Smith [18, 19] demonstrate the systematic development of an IQ model for Wikipedia. They start from a synopsis of the Wikipedia quality model, the Crawford model and a model that the authors have developed for the quality assessment of Dublin core metadata records. Additionally Wikipedia discussion pages are evaluated. The quality issues mentioned there are integrated into the model. As a next step the quality dimensions of the model are mapped to a metric by applying factor analysis to samples of random articles and featured articles. This metric is based on text properties – like text length, number of links, edits or authors etc. – that can be ascertained automatically in an efficient way. Merits of this approach are the strict method of model construction and the ease of application with respect to the resulting metric. However, there are some shortcomings also. The metric – not the general model – is based on the vote of one of Wikipedia's communities only. It is questionable, whether the model is applicable to the German or French part of Wikipedia. The model aims primarily at article quality. Quality criteria – like coverage – which apply to the encyclopedia as a whole are neglected. Some aspects of hypertext structure – internal, external and broken links – are taken into consideration others aren't – interlanguage links, categories. A motivation for this selection is not provided. The community seems to prefer long articles. Thus, the model cannot develop a notion of an appropriate amount of information, which includes an upper bound for text length.

The models described above provide a sound basis for further investigations on Wikipedia quality. The study presented in this paper aims at filling some of the gaps left.

1.3 *Studies on Wikipedia`s information quality*

Some studies on Wikipedia's information quality are already available. A few of them aim primarily at the development of a quality model – the studies by Stvilia et al. belong to this group – others find their result without bothering about a reusable model.

Lih [14] presents a simple quality metric based on the number of edits and authors of evaluated articles. This study was the first one to use quantitative methods in the context of Wikipedia evaluation. Emigh and Herring [2] evaluate the language register of Wikipedia articles as an important formal aspect of information quality.

There are some comparative studies as well. The *Nature* study [4] compares Wikipedia and Britannica on the base of a small expert evaluation. This study aims at factual accuracy only, where the studies of Schlieker [17] and Hammwöhner et al. [6] compare Wikipedia and the Brockhaus encyclopedia according to intrinsic information quality criteria like coverage, text length, linking, verifiability etc. It is worth mentioning that Wikipedia improved its position in comparison to Brockhaus in the time between these two studies.

None of the studies above deals with knowledge organization or software usability. Only newer studies [23, 7] take this quality dimension into consideration. Wiegand [23] presents an expert evaluation, whereas Hammwöhner [7] offers a user test additionally. The results are once more quite positive. The user test, however, showed some significant shortcomings with respect to knowledge organization and interface. Most of the participants, for instance, didn't even know about the category system and, thus, were not able to use it for information search.

1.4 *What has to be done?*

There are substantial research results about Wikipedia at hand comprising a well founded quality model and lots of empirical findings. There are, however, some gaps. One would like to know whether the model as defined by Stvilia et al. [18,19] is valid not only for the English Wikipedia. Further investigations seem to be necessary as far as the category system is concerned. No quality metric is available here, but traditional models of indexing quality [16] can be adopted. Available approaches to quality assessment for Wikipedia are able to compare or rank resources according to their quality. They can classifiy quality problems according to quality dimensions. But they will not point to concrete instances of quality problems. In the following we will investigate whether the exploitation of interlanguage redundancy may help to deal with some of these issues.

1.5 *Relevance of this research*

The scope of research presented in this paper seems to be very limited since it focuses on interlanguage linking and categorization within Wikipedia only. These aspects, however, are of relevance to a comparatively wide field of applications. An obvious consequence of improvements concerning the category system and the indexing process will be an increase in retrieval quality. Wikipedia's category system is not only an instrument of information retrieval but serves as a basic structure for text mining processes as well. Ponzetto and Strube [15], for instance, use the category system of the English Wikipedia to construct a large scale taxonomy, while Völkel et al. [21] propose to built semantic representations from Wikipedia structures – e.g. the category system and the hyperlink-graph. The quality of the category system therefore is of crucial importance for the quality of the resulting information services. Awareness of quality problems, thus, can be regarded as a first premise of successfully use Wikipedia as a subject of semantic interpretation [8]. In our approach we use interlanguage redundancy firstly for the detection of quality problems and in a second step for the derivation of improved representation structures. It may be seen as complementary to approach developed by Ponzetto and Strube [15], which is based on mono-lingual evidence only. The comparison and integration of structures stemming from various language specific category systems is a task similar to ontology alignment [11]. Methods from this research field will be used within the further course of our project. Modifications to the standard procedures of ontology alignment will be necessary because of the size of Wikipedia's category system and because of the nature of the weakly structured common sense knowledge it represents. Any findings in that area will be of relevance for ontology alignment in general.

Some publications – e.g. [12] – try to apply theories of crowd intelligence [20] to Wikipedia. Phenomena of crowd intelligence can be found if and only if the following conditions are fulfilled:

- **Diversity of opinions** seems to be guaranteed by the large community of Wikipedia authors.
- **Independence of judgments** may suffer from communication between authors via the talk page. The authority of administrators or power users may additionally affect the independence of judgment. Kittur et al. [12] point out that this effect is of decreasing significance at least as far as the English Wikipedia is concerned.
- **Decentralized organization** is guaranteed by the loosely coupled communication structure of Wikipedia.
- **Existence of an algorithm for the integration of judgments** – Wikipedia provides support for the replacement but not the integration of versions.

Only the fourth criterion seems to impose problems. This can be shown most clearly in the context of the category system. Category assignment is a binary choice. Most cooperative tagging systems [5] allow multiple category assignments, which are then weighted according to their frequency. This approach may be conceived as some kind of integration of judgments. A similar effect may be achieved for Wikipedia by the integration of category assignments made in different languages.

Research about category assignment in Wikipedia provides some insight into the conceptualization of common sense knowledge, but the main achievement will, hopefully, be the improvement of cooperative information systems.

2. GENERAL APPROACH

The general goal of the project is the search for information patterns within Wikipedia – or other Web 2.0 applications – that may be found by means of bibliometrics or webometrics. In the scope of this paper the focus will be on interlanguage differences concerning quality features and categorization. The research presented here is still in an early stage. Thus, the focus is on hypothesis generation not on hypotheses testing. The data presented are extracted from the English (*en*), German (*de*), French (*fr*), Italian (it), part of Wikipedia.

The data extraction and evaluation was performed by a tool developed by the author. It is capable of:

- Downloading samples of Wikipedia articles, which are selected from previously defined lists (e.g. excellent articles), at random using Wikipedias random function, by random walk starting from a pre-given seed, or by crawling the Wikipedia web.
- The accessed Wikipedia articles may be evaluated with respect to text length (no. of words), link density, currency, no. of versions and authors, categories etc. These data may similarly be obtained from talk pages as well. The data available are compatible to the quality dimensions as proposed by Stvilia et al. [19]. As the only exception readability indices are not used, since they may not be used for interlanguage comparison.
- The tool can process data from the Catalan, Czech, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian (Norsk and Nynorsk), Polish, Portuguese, Spanish, Swedish and Turkish part of Wikipedia. Adding an additional language implies approximately 1 hour of work – notwithstanding that some evaluations like word counting can be used with alphabetic writing systems only.

The data, which will be presented in this paper, are chosen to support the following course of argumentations:

- The first sample (section three) will show general features of the four selected Wikipedias with respect to standard and featured articles.
- Section four deals with the reliability of interlanguage links, which are used for the further evaluation processes.
- Section five deals with quality issues concerning the structure of the category system (5.1) and the use of categories as index terms (5.2).
- Section six shows some simple heuristics, which improve the structure of the category system based on interlanguage links.

3. COMPARING: EN, DE, FR, IT

This small study is based on downloads (1st week of May) of *all* featured articles (en: 1364, de: 1044, fr: 350, it: 299) and random samples of 1100 articles (after removal of redirect etc.: en: 1059 de: 989, fr: 1032 it: 1055). The total size of Wikipedia at that time was: *en*: 1.763.740, *de*: 577.920, *fr*: 483.875, *it*: 290.684 [30]. Table 1 shows, that the differences between standard and featured articles are by far greater that those between languages. This effect is valid not only for the depicted quality criteria, but for all features mentioned by Stvilia et al. [19].

	categories	out-links	in-links	external-links	text-length
en	3 (22)	19 (762)	6 (1.820)	1 (77)	271 (8.676)
en featured	6 (42)	203 (936)	129 (2.459)	24 (215)	5744 (22.732)
de	2 (15)	20 (1.121)	7 (1.007)	1 (312)	309 (12.562)
de featured	4 (20)	135 (1.011)	55 (1.723)	7 (312)	4627 (29.047)
fr	2 (16)	21 (412)	6 (999)	1 (48)	270 (11.512)
fr featured	4 (15)	187 (890)	70 (1.477)	11 (118)	6392 (33.098)
it	2 (9)	24 (381)	5 (1.000)	1 (78)	283 (6910)
it featured	4 (22)	191 (1055)	112 (2.480)	6 (98)	5144 (28.884)

Table 1: Selected quality featured (median, maximum values in brackets)

The data suggest, that the process of model construction as proposed by Stvilia et al [19] can by applied to other languages as well and should lead to similar results. The weighting of individual features within the quality metric, however, will presumably be language specific. Detailed statistical analysis has to confirm this assumption.

4. QUALITY OF INTERLANGUAGE LINKS

Interlanguage links will play a major role in the evaluation of the category systems as presented in the next chapter. This requires a pilot study that deals with the quality of this representation mechanism itself. The following cases of interlanguage link assignment may be distinguished:

a. Interlanguage links are assigned consistently and correctly between two articles,
b. Interlanguage links are missing,
c. Links are assigned erroneously in both directions,
d. Links are assigned inconsistently: one article links to some other, but no link is set in the reverse direction.

Obviously, only case *d* can be identified automatically. Case *c* seems – detailed analysis still has to be done – not to occur at a significant rate. The missing of interlanguage links can be observed quite often. Instruments for the (semi-) automatic detection of this error can be developed in combination with additional tools only (see below).

Inconsistent linkage may result from a user simply forgetting to define the inverse link. This kind of error can be detected and corrected automatically. Wikipedia already employs automatic agents for this purpose. In other cases inconsistent interlanguage links indicate inconsistencies in the modeling of lemmata. The following phenomena – explained by examples – may occur:

- One of the links erroneously point to a disambiguation page: German *Main*, links to the Italian disambiguation page *Meno*.
- One of the links points to a wrong translation: a German page about the American comic book *Preacher* links to the lemma about preacher in the literal sense of the word in several languages (Italian, French, Swedish etc.).
- There are lemmata with redundant or overlapping content in one of the languages, which, perhaps, should be integrated: Both of the English lemmata *Figure of speech* and *Rhetorical device* link to the German article *Rhetorische Figur*, which links back to *Rhetorical device* only.
- A complex phenomenon is described by one lemma in the first language and by two or more lemmas in the second one. This may not be considered as an error but as a matter of design and style. The English Wikipedia presents information about the proteins *fibrin* and *fibrinogen* in a single article, whereas its German counterpart has two articles on these topics.

Up to 5% of the interlanguage links from various bilingual samples (de-en, de-fr, de-it) were affected by inconsistencies, which could be detected and corrected easily.

languages	median	maximum
en	0	61
en featured	10	191
en categories	2	123
de	1	114
de featured	11	179
de categories	11	124
fr	1	97
fr featured	16	154
fr categories	7	122
it	3	78
it featured	21	153
it categories	6	122

Table 2: Number of interlanguage links (same sample as in table 1)

The presence of interlanguage links seems to be an indicator of information quality, since featured articles have significantly (α=0.01) more interlanguage links than normal ones. This is probably due to the fact, that good articles are more likely to be translated. Categories (sample described in detail in section 5.1) also have a comparatively high density of interlanguage links.

5. CATEGORIES IN: *EN, DE, FR, IT*

The category systems – each language has its own – of Wikipedia can be understood as some kind of controlled and structured vocabularies. Wikipedia articles may be assigned to an arbitrary number of categories. Categories may be categorized themselves. Thus, a category system is a directed graph of categories connected by links between subcategories and categories. Additional links to related terms may be inserted within the category definition. Corresponding categories can be connected by interlanguage links just like normal articles. The resulting structures are quite similar to those of traditional thesauri. This interpretation – as Voss points out [22] – is supported by an analysis of Wikipedia's category system. The practical use of the category system, however, is quite different from the use of thesauri. Links leading to related terms are used in rare cases only. Cycles in the category structure occur.
A structured category system like Wikipedia's could be useful for several purposes:

- The quality of Wikipedia's search mechanisms could be improved. Some search tools that make use of the category systems exist, but they are not integrated into the standard user interface.
- Categories may be used as an alternate means of navigation. The user may switch from the content name space to the category space and explore articles, which are assigned to the same category. He may browse the category structure in order to find articles about more general or special subject fields. According to Wikipedia's internal documents [29] this kind of category use seems to be the canonical one.
- Categories of an adequate abstraction level could be used to group thematic clusters within the encyclopedia such that the thematic coverage can be evaluated. If the category systems of Wikipedia's languages were compatible, a comparative evaluation of thematic coverage would be possible.
- The category system may be used for text mining [15] and semantic interpretation of Wikipedia texts [21].

In the following we will check the suitability of Wikipedia's category systems for the above mentioned tasks. In a first step we will check information quality features of the category systems. A second step will be the assessment of category assignments.

5.1 Information quality of Wikipedia's category systems

The following dimensions of information quality are proposed for indexing systems [3].

- High *coverage* indicates, that all relevant thematic fields can be described by terms of the controlled vocabulary.
- *Precision* is about the degree of detail that can be covered by the indexing system. Since the category system of Wikipedia is open to expansion on demand, this criterion is not applicable.
- *Consistency* is granted, if no unwanted or contradictory inferences can be made from the category system.
- *Expandability* refers to the possibility to enhance the vocabulary. Since Wikipedia offers the option to enter new categories on demand, this criterion seems to be met. The dependency between expandability, usability and consistency hast nevertheless to be discussed.
- *Usability*: Indexing processes can be performed with reliable result only, if sufficiently specific comments define the scope of categories. Furthermore the overall structure of the vocabulary must be manageable and comprehensible.
- Economic viability: The development of controlled and structured vocabularies normally is a costly process. An appropriate notion of costs in the context of social computing, however, has still to be developed. We will not take this quality dimension into account within the scope of this paper.

The study presented here will not give full account of all of the quality dimensions mentioned above, but aims at special aspects of consistency and usability. The question to be answered is: does cooperative construction of a controlled and structured vocabulary result in manageable and consistent structures. The evaluation of indexing systems with respect to usability would ask for user testing as presented in [7]. Since cooperative tagging is not well understood up to now, it is advisable to start with less costly studies to get an impression of the structures which evolve from cooperative indexing. This first study is based on a multi lingual sample from the English, French, German and Italian Wikipedia. 463 Italian articles were selected randomly such that translations to English, French and German exist. Table 3 shows how many descriptors are employed to describe the articles from these samples. Depth and size of the category tree built upon these basic categories indicate the complexity of the overall structure. Additionally the number of cycles to be observed in the category tree has been counted. Obviously the structures of the category systems of the four Wikipedias concerned differ by far. Two of them – *en* and to a minor degree *fr* – employ a very rich and complex structured category system, whereas the others rely on a smaller number of categories ordered within a by far simpler structure. It can be assumed, that any user will have difficulties to memorize a multi-hierarchy with a depth (maximum of the shortest path length to the top node) of 11 or even 15. But if not only shortest paths but also longest paths are taken into consideration, the differences become more obvious. Inconsistencies within the category structure have to be expected as a result of this complex structure. The task of choosing categories for indexing is very demanding, too. Inconsistent indexing is to be expected as well. It is worth mentioning, that size and complexity of the category system has no direct correlation to the size of the Wikipedia, since *en* and *de* are the largest Wikipedias, whereas *en* and *fr* employ the most complex category system.

	cat. / lemma (median)	Subcat. / cat (median)	descriptors	total no. of cat.	depth	longest path length	no. of cycles
en	4	7	1878	11916	12	138	432
de	2	5	1048	3422	12	18	8
fr	4	5	1173	4504	15	63	66
it	2	4	800	2191	11	18	13

Table 3: Use of categories in multilingual samples

Indeed, semantically inconsistent structures can be found easily within the category systems. We will have a look at a comprehensive example which has been picked from the category system of the English Wikipedia by chance. '→' denotes a link between a category and one of its supercategories, the numbers

in brackets denote:

- the number of Wikipedia languages in which this term is present besides English,
- the number of subcategories,
- the number of entries within this category.

This example will show us the interrelations between Virgil and Mesozoic animals, reality and advertising, literature and physical quantities, or between almost everything.

Aeneid (0/1/8) → Poetry of Virgil (0/1/4)→ Poems by author (0/32/0) → Works by author (1/144/0) → Literature (86/64/109) → Arts (72/18/35) → Aesthetics (15/12/68) → Perception (19/13/111) → Psychology (69/37/>200) → Behavioural sciences (5/6/34) → Behavior (17/11/54) → Nature (45(24/9) → Knowledge (18/20/55) → Information (26/13/39) → Physical quantity (35/19/138) → Physics (88/36/183) → Science (89/51/134) → Academic disciplines (11/11/27) → Academia (17/25/174) → Education (53/45/5) → Personal development (4/10/102) → Social psychology (19/22/>200) → Crowd psychology (0/8/25) → Public opinion (0/10/25) → Group processes (0/4/68) → Anticipatory thinking (0/5/49) → Futurology (19/24/132) → Future (9/11/10) → Time (60/13/60) → Metaphysics (23/19/92) → Reality (0/8/14) → Philosophical concepts (13/7/167) → Philosophical terminology (18/0/>200) → Vocabulary (4/9/29) → Language (5/14/20) → Human communication (10/15/76) → Communication design (2/9/117) → Advertising campaigns (1/4/63) → Advertising (23/25/196) → Media by interest (0/19/1) → Mass media (44/35/87) → Information science (13/17/100) → Applied sciences (34/18/23) → Technology (59/35/138) → Humans (24/20/27) → Apes (16/6/37) → Primates (36/9/7) → Mammals (66/35/46) → Cynodonts (1/1/33) → Mesozoic animals (2/8/2) → Mesozoic life (0/1/3) → Prehistoric life (2/13/5) → Prehistory (23/12/79) → Anthropology (63/40/>200)

It can be learned from this example, that the category-subcategory-relation of Wikipedia's category systems may not be confused with some kind of is-a-relation. The link between a category and its subcategories may express an is-a-, a part-of-, or simply a related-term-relation. This means, that transitive links between categories have no reasonable interpretation at all. This finding should be taken into account, when approaches to the semantic interpretation of Wikipedia structures are proposed [21].

The evaluation shows, that the German and the Italian category system have some advantages (smaller, less complex, small amount of cycles) with respect to usability. Inconsistencies do not show up as obviously as in the example above. More research in this area, however, has to be done. The English and French category systems, on the other hand, cover by far more details and thus allow for a higher precision of indexing. Whether a distinct category *Poetry of Virgil* or even more special *Aeneid* (see above) is truly needed, can be confirmed finally by user tests only. Nevertheless, the number of languages specific category systems containing some specific category may serve as an indicator of the need for its existence.

5.2 Information quality of category assignment

The assignment of categories to Wikipedia articles can be evaluated according to various criteria:

- Consistency of indexing within a Wikipedia: Are similar articles described similarly? This criterion is of great importance if a category system is to be used for the purpose of query oriented information retrieval. We will not pursue this question within this paper.
- Consistency between Wikipedias: Are corresponding articles, which are connected consistently by interlanguage links, indexed in a consistent way? If this is the case, interlanguage query support is feasible. Furthermore comparisons between Wikipedias according to coverage could be possible without much effort.
- Population: The utility of category systems depends on the population of the categories. Weakly populated categories make a strong distinction. They are not useful to identify thematic clusters within a text sample. They aren't appropriate for query support as well, since only very specific

questions can be answered. But they can be explored easily by navigation. Highly populated categories on the other hand don't make a sufficient distinction. They can not be used properly for querying or navigation but for large scale thematic analysis. A look at the statistics pages of the Wikipedias in question shows, that there is an enormous statistical dispersion in the population of categories. There are dozens of categories containing thousands or even tens or hundrets of thousands of articles and lots of categories containing no more than up to ten articles. The implication is, that no general advice for an appropriate use of the category system can be given. Proper use depends on the proper choice of categories. Conventional information retrieval systems employing structured vocabularies would provide the option to switch from special to general categories. This is not possible within Wikipedia since the relation between categories and subcategories is not transitive (see above).

In the following we will report on an experiment concerned with indexing consistency between Wikipedias. The samples of table 3 are reused. For every pair of corresponding articles some measures of interlanguage indexing consistency are computed. They were derived from the Jaccard similarity coefficient, a standard measure for the similarity of sample sets [16]. If A is a first set of index terms and B a second one, then the similarity between these sets is defined as:

$$J(A,B) := \frac{|A \cap B|}{|A \cup B|}$$

This measure is not directly applicable, since index terms of two different languages are concerned. If $trans_{c1,c2}$ is a mapping from a set of categories contained in category system 1 to a set of categories contained in system 2 (if available), then the measure is:

$$IC1_{c1,c2}(A_{c1}, B_{c2}) := \frac{|trans_{c1,c2}(A_{c1}) \cap B_{c2}|}{|A| + |B| - |trans_{c1,c2}(A_{c1}) \cap B_{c2}|}$$

This measures relates the number of corresponding index terms to the total number of index terms. A second measure relates the number of corresponding index terms to those, which have a translation.

$$IC2_{c1,c2}(A_{c1}, B_{c2}) := \frac{|trans_{c1,c2}(A_{c1}) \cap B_{c2}|}{|trans_{c1,c2}(A)| + |trans_{c2,c1}(B)| - |trans_{c1,c2}(A_{c1}) \cap B_{c2}|}$$

IC 3 and IC 4 were derived from IC 1 and IC 2 respectively by allowing a category to match with one of its subcategories. Table 4 shows that interlanguage indexing consistency is very low in general. It is not sufficient for interlanguage information retrieval. There is, however, one promising effect. Up- or down-posting by one level improves the results by far. Some of the bad results can be explained by missing interlanguage links. The evaluation tool will produce a list of possible interlingual links as a first step to improve the knowledge organization structure.

One could assume that all articles showing a good interlanguage indexing consistency have a high percentage of international authors. Indeed there are some articles, which seem to show that effect, but detailed statistical analysis has still to be done.

	IC 1	IC 2	IC 3	IC 4	Shared authors
en-de	0.14 (0.5)	0.33 (1.0)	0.38 (0.5)	0.71 (1.0)	0.02 (0.25)
de-it	0 (0.5)	0 (1.0)	0.13 (0.5)	0.6 (1.0)	0.05 (0.33)
de-fr	0.13 (0.5)	0.33 (1.0)	0.38 (0.5)	0.73 (1.0)	0.06 (0.33)
fr-it	0 (0.5)	0 (1.0)	0.17 (0.5)	0.5 (1.0)	0.06 (0.33)

Table 4: Indexing consistency measures (median and maximum) and percentage of shared registered authors (by name)

6. EVALUATION OF MULTILINGUAL CATEGORY SYSTEMS

Section 5.1 demonstrated quality problems within Wikipedia's category system based on a substantial example. In this section we will discuss some simple heuristics, which may help to detect and, in some cases, to solve these problems. The following types of problem may be discerned easily:

- Lack of distinction between subcategories and related terms,
- confusion about the direction of the sub-category relation,
- inappropriate level of detail within the structure of the category-system.

The first two kinds of problems may lead to erroneous interpretations. A lack of detail within the representation will cause bad retrieval quality, where category systems with a high amount of details require an excessive maintenance effort.

A first, crude heuristic would remove all categories, which are not shared by a certain number of Wikipedias. A look at our example reveals that most of the problems would cease to exist. Since the English Wikipedia is by far larger than the others, some of these categories may be necessary to represent its additional amount of content. Categories of this type should contain a minimal number of subcategories or entries. This is probably the case with *public opinion* but not with *poems of Virgil*.

The interlingual connection of two categories does not imply that they share corresponding sub- or super-categories. A good match between the links structures of different category systems seems to be a good indicator for semantic validity. To evaluate this heuristic we have checked our example with the category systems of the Czech, Danish, German, Finnish, French, Italian, Dutch, Norwegian (Bokmål), Polish, Portuguese, Swedish and Turkish Wikipedias. Only some few category pairs got a confirmation of more than 80% (Physical quantity → Physics → Science, Primates → Mammals). Some more reached a rating of about 50% (Literature → Arts, Aesthetics → Arts, Technology → Applied Sciences). It is worth mentioning that the majority of Wikipedias considers – in contrast to the English version – *Technology* as subordinate to *Applied Sciences* and *Aesthetics* as subordinate to *Arts*. 50% of the Wikipedias accept *Advertising* as directly subordinate to *Mass media*. All of the Wikipedias taken into account ignore *Media by interest*. This is a structural indicator for this category being of some special type. It is not used to describe contents but to organize the category system.

The examples suggest that the simple heuristics introduced here are capable to detect and solve some of the quality problems that arise within Wikipedia's category systems. This finding, however, has to be confirmed by a comprehensive survey. As a further step more subtle heuristics have to be developed and integrated within a formal model of ontology or thesaurus alignment. It seems to be most promising to adapt a model of fuzzy thesauri as proposed by Intan and Mukaidono [10] to the specific needs of Wikipedia. The respective fuzzy values have to be derived from the heuristics.

7. CONCLUDING REMARKS

The paper reports on quality issues in the scope of Wikipedia that have not been dealt with in previous research. Fragments of an extended model for information quality within cooperatively edited texts show up. But further research is needed here. The proposed approach of interlingual quality assessment allows, as has been demonstrated, for the detection of specific quality problems – missing or ill structured categories, missing interlanguag links etc. Quite similar tests can be applied to the link graph of Wikipedia. The main goal of research will be the further development of models and tools for the quality assessment of global information structures as observed in Wikipedia and equivalent electronic encyclopedias. If this research would give an explanation for the apparent differences in the development of the category systems of the English and French Wikipedias on one side and the Italian and German ones on the other, this would be a great asset since a better understanding of cooperative processes would be achieved.

REFERENCES

[1] Crawford, H. Encyclopedias. R. Bopp, L.C. Smith (Eds.) *Reference and information services: an*

introduction. Libraries Unlimited. 2001. pp. 433-459

[2] Emigh, W., Herring, S.C. Collaborative authoring on the web: A genre analysis of online encyclopedias. *Proc. Annual Hawaii International Conference on System Sciences*. 2005. http://ella.slis.indiana.edu/~herring/wiki.pdf. cited at 1/6/2007

[3] Fettke, P., Loos, P. Komponentendokumentation – Eine systematische Bewertung von Ordnungssystemen aus formaler Sicht. *Proc. MobIS 2000*. pp. 51-70. http://archiv.tu-chemnitz.de/pub/2001/0025. cited at 7/3/2007

[4] Giles, J. Internet encyclopaedias go head to head. *Nature* 438. 2005. S. 900-901, http://www.nature.com/nature/journal/v438/n7070/full/438900a.html. cited at 1/6/2007

[5] Golder, S., Huberman, B. A. The structure of collaborative tagging systems. *Journal of Information Science*. 32(2). 2006. pp. 198-208. http://arxiv.org/abs/cs/0508082v1, zitiert am 22.8.2007

[6] Hammwöhner, R., Fuchs, K.-P., Kattenbeck, M., Sax, C. Qualität der Wikipedia. Eine vergleichende Studie. Oßwald, A., Stempfhuber, M., Wolff; C. (eds.) *Open Innovation. Proc. 10th Int. Symposium on Information Science*. UVK, 2007, pp. 77-90

[7] Hammwöhner, R. Qualitätsaspekte der Wikipedia. Stegbauer, C., Schmidt, J., Schönberger, K. (eds.) Wikis: Diskurse, Theorien und Anwendungen. *kommunikation@gesellschaft*. Special issue on Wikis. 8, 2007. http://www.soz.uni-frankfurt.de/K.G/B3_2007_Hammwoehner.pdf. cited 16/9/2007

[8] Hammwöhner, R. Semantic Wikipedia – checking the premises. Accepted for the SABRE conference on social semantic web. 2007

[9] Hammwöhner, R. Wikipedia – ein Medium der Ignoranz? Geisenhanslüke, A. (ed.) Ignoranz. Transcript. 2007. To appear. Preprint: http://www-nw.uni-regensburg.de/%7E.har16557.infwiss.sprachlit.uni-regensburg.de/ Literatur/ignoranz_2007.pdf. cited at 13/9/2007

[10] Intan, R., Mukaidono, M. Generating fuzzy thesaurus by degree of similarity in fuzzy covering. Zhong, N., Ras, Z.W., Tsumoto, S., Suzuki, E. (eds.) *Foundations of intelligent systems*. Lecture notes in computer science. Springer. 2003. pp. 427-432. http://www.springerlink.com/content/hru3tpgw8wql1a3k/. cited 15/9/2007

[11] Kalfoglou, Y.; Schorlemmer, M. Ontology mapping: the state of the art. *The Knowledge Engineering Review*,.18 (1). 2003. pp. 1–31.

[12] Kittur, A., Chi, E., Pendleton, B. A., Suh, B., Mytkowitz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007); 2007 April 28 - May 3; San Jose; CA. 2007

[13] Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y. AIMQ: a methodology for information quality assessment. *Information and Management*. 40 (2). 2002. pp. 133-146

[14] Lih, A. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *Proc. of the 5th International Symposium on Online Journalism*. 2004. http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf. cited at 1/6/2007

[15] Ponzetto, S. P., Strube, M. Deriving a large scale taxonomy from Wikipedia. *Proc. 22nd National Conference on Artificial Intelligence*. 2007. pp. 1440-1445. http://www.eml-research.de/english/homes/ponzetto/pubs/aaai07.pdf. cited at 13/9/2007

[16] Salton, G., McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company. 1984

[17] Schlieker, C. *Explorative Untersuchung von Wissen in kollektiven Hypertexten*, Diplomarbeit, Fachbereich 08, Soziologie, Universität Bremen, 2005

[18] Stvilia, B., Twidale, M. B., Gasser, L., Smith, L. C. Information Quality Discussions in Wikipedia. In: S. Hawamdeh (Ed.), *Knowledge Management: Nurturing Culture, Innovation, and Technology - Proceedings of the 2005 International Conference on Knowledge Management*. Charlotte, NC: World Scientific Publishing Company, 2005, pp. 101-113. http://mailer.fsu.edu/~bstvilia/papers/qualWiki.pdf. cited at 1/6/2007

[19] Stvilia, B., Twidale, M. B., Gasser, L., Smith, L. C. Assessing information quality of a community-based encyclopedia. In: *Proceedings of the International Conference on Information Quality - ICIQ 2005*. Cambridge, MA, 2005, S. 442-454. http://mailer.fsu.edu/~bstvilia/papers/quantWiki.pdf. cited at 1/6/2007

[20] Surowiecki, J. *The Wisdom of the Crowds. Why the Many are Smarten than the Few and How Collective Wisdom Shapes Business, Economics, Societies, and Nations*. Doubleday. 2004

[21] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R. Semantic Wikipedia. In *Proc. 15th Int. Conf. on World Wide Web, WWW 2006*, Edinburgh, Scotland, May 23-26, 2006. http://www.aifb.uni-karlsruhe.de/WBS/hha/papers/SemanticWikipedia.pdf, cited at 5/20/2007

[22] Voss, J. Collaborative thesaurus tagging the Wikipedia way. *Wikimetrics research papers*. 1 (1). http://arxiv.org/abs/cs.IR/0604036. cited at 6/7/2007

[23] Wiegand, D., Entdeckungsreise, Digitale Enzyklopädien erklären die Welt, *c't, Magazin für Computer und Technik*, Nr. 6, 2007, S. 136-145
[24] Wikipedia: *About*. http://en.wikipedia.org/wiki/Wikipedia:About. Cited at 12/9/2007
[25] Wikipedia: *Featured article criteria*. http://en.wikipedia.org/wiki/Wikipedia:What_is_a_featured_article. cited at 6/30/2007
[26] Wikipedia: *Featured picture criteria*. http://en.wikipedia.org/wiki/Wikipedia:Featured_picture_criteria. cited at 6/30/2007
[27] Wikipedia: *Featured list criteria*. http://en.wikipedia.org/wiki/Wikipedia:What_is_a_featured_list. cited at 6/30/2007
[28] Wikipedia: *Featured portal criteria*. http://en.wikipedia.org/wiki/Wikipedia:What_is_a_featured_portal. cited at 6/30/2007
[29] Wikipedia: *Categorization*. http://en.wikipedia.org/wiki/Wikipedia:Categorization, cited at 6/30/2007.
[30] Wikipedia: *Multilingual ranking May 2007*. http://en.wikipedia.org/wiki/Wikipedia:Multilingual_ranking_May_2007. cited at 7/2/2007