Tagging Tagging.
Analysing User Keywords in Scientific Bibliography Management Systems

Markus Heckner (markus.heckner@paedagogik.uni-regensburg.de), Media Educational Science
Susanne Mühlbacher (susanne1.muehlbacher@sprachlit.uni-regensburg.de), Information Science
Christian Wolff (christian.wolff@computer.org), Media Computing
University of Regensburg
93040 Regensburg
Germany

Abstract
In this paper, an empirical study of tagging behaviour in web-based bibliographic annotation systems is presented. Starting from an initial category finding phase in which tags attributed to selected articles from *Connotea* were classified we have set up a category model for linguistic and functional aspects of tag usage as well as for the relationship between tags and document full text. In a second phase this model is applied to approx. 500 tagged articles from the information and computer technology domain randomly selected from *Connotea*. Our findings show significant differences to other tagging research which was primarily conducted using popular (non-scientific) tagging platforms like *Flickr* or *Delicious*. We observe a great overlap of tag material and document text and rather few non-content related tags. The comparison of user tags with author keywords shows that users tend to use less and more general tags. Finally, system functionality seems to play a role for users' tagging behaviour.

1 Research Context and previous work

Recently, a growing amount of systems that allow content annotation by their users (i.e., *tagging*) has been created, ranging from personal sites for organising bookmarks (*http://del.icio.us*), photos (*http://flickr.com*) or videos (*http://video.google.com*, *http://youtube.com*) to systems for managing bibliographies for scientific research (*http://citeulike.org*, *http://connotea.org*). Simultaneously, a debate on the pros and cons of allowing users to add personal keywords to digital content has arisen (e.g. Shirky 2005a).

One recurrent point of discussion is whether tagging can solve the well-known vocabulary problem: In order to support successful retrieval in complex environments, it is necessary to index an object with a variety of aliases (cf. Furnas et al. 1987). A thesaurus may assist users in achieving a better match between their search query and the indexing terms provided by indexers and authors (Foskett 1997, Lancaster 1993) by presenting concepts related to their search terms. In this spirit, social tagging enhances the possibilities of traditional (author or expert) indexing by adding user-created retrieval vocabularies which could bridge the gap between users and authors or indexers without the expensive process of creating a thesaurus or cross-vocabulary concordance.

Furthermore, tagging can go beyond *content-related* keywords by providing meta-keywords like *funny* or *interesting* that "identify qualities or characteristics" beyond mere content description (Golder and Huberman 2006, Kipp and Campbell 2006, Kipp 2007, Feinberg 2006, Kroski 2005). On the contrary, tagging systems are claimed to lead to semantic difficulties that may hinder the precision and recall of tagging systems (e.g. the polysemy problem, cf. Marlow 2006, Lakoff 2005, Golder

and Huberman 2006, Shirky 2005b). These problems have been recognized and some attempts to structure the tag space from rather different angles have been made: Xu et al. (2006) propose a set of general criteria for a successful tagging system, while Schmitz (2006) attempts to induce vocabulary from *Flickr* tags. On the other hand, Begelman et al. (2006) report improvements in searching and navigating the tag space by adding clustering techniques to a tagging system. Finally, Aurnhammer et al. (2006) propose a combination of emergent semantics and tagging by using visual features to help users discover new relationships between data.

Empirical research on social tagging that goes beyond implementing and evaluating individual systems built for a specific purpose is still rare. Some case studies which employ tagging to solve problems in an enterprise scenario are available. For example, Farrell and Lau (2006) extend the list of "taggable" resource types by tagging people in order to improve contact organization and to inform users of other peoples' skills and expertise. In John and Seligman (2006) the potential of tagging in the enterprise is discussed and an approach to rank experts based on tagging activity is presented. Damianos (2006) explores the potential benefit of tagging in a corporation (cf. Dennis (2006), Trant and Wyman (2006) and Bar-Ilan et al. (2006) for further case studies).

The larger part of remaining research approaches comes from a computer linguistics or librarian point-of-view (Voß 2007) and focuses either on the automatic statistical analyses of large data sets, or intellectually inspects single cases of tag usage: Some authors studied the evolution of tag vocabularies and tag distribution in specific systems and contexts (Golder and Huberman 2006, Hammond 2005, Yew and Teasley 2006). Others concentrate on tagging behaviour and "tagger" characteristics in collaborative systems (Hammond 2005, Kipp 2007, Feinberg 2006, Sen 2006).

However, little research has been conducted on the functional and linguistic characteristics of tags (initial ideas and findings can be found in Kipp and Campbell 2006, Kipp 2007, Golder and Huberman 2006 which serve as a starting point for our study). Analysing these patterns could show differences between user wording and conventional author- or expert-based keywording. In order to provide a reasonable basis for comparison, a category model for existing tags is needed. Additionally, most research seems to have focussed on systems for private or personal use like *Delicious* or *Flickr*. In this paper, we analyse tag usage in *Connotea,* a system for the management and sharing of scientific bibliographies. Our main research questions can be stated as follows:

- Is it possible to discover regular patterns in tag usage and to establish a stable category model for tags?
- To what degree are social tags taken from or findable in the full text of the tagged resource?
- How do social tags differ from author keywords?

- Does a specific tagging language comparable to internet slang or chatspeak evolve?
- Do tags in a research literature context go beyond content description (e.g. tags indicating time or task-related information, cf. Kipp and Campbell 2006)?

2 Goals and Methodology

Our study has two major goals: The creation of a category model for tags and an empirical analysis of tag usage in the *Connotea* context using this model for tag classification.

2.1 Dataset and Tag Category Model (TCM)

Our study was conducted in two steps using data from *Connotea*:

Step 1: Explorative creation of a tag category model
By utilizing *Connotea's* web API all posts that were uploaded or added to the *Connotea* database on 6 November 2006 were retrieved in a single XML document. Each post contains the title, the source and the tags that were assigned by the uploader[i]. The XML format was subsequently transformed into more accessible HTML and distributed among four information scientists. This group of experts consisted of two PhD students in Information Science and two professors of Information Science and Media Computing. The instruction for these experts was to try to derive possible reccurring patterns in tag usage from the data. The experts had access to the following information: resource name, URL to the post in *Connotea* and the tag itself. Each expert developped suggestions for possible functional as well as linguistic tag categories. Following this individual analysis an expert workshop was conducted in order to integrate all individual suggestions into a preliminary category model. In this workshop all category suggestions were discussed and weighted, differing labels mapped onto preferred category names and suggestions with little data in support of them discarded.

Step 2: Explanatory case study: Applying and verifying the category model.
An additional data set of 500 information and computer technology (ICT)-related scientific articles was extracted from *Connotea.org*. These randomly selected articles had to match two criteria: Access to the document's full text as well as author keywords had to be available. The same expert group as in step 1 (familiar with ICT topics) was instructed to introspectively assign the 1191 user tags as well as the dodument's author keywords to the category model in order to define and compare functional and linguistic characteristics. A main goal of the second step was to verify the preliminary category model: For this purpose the data was categorised using Excel spreadsheets that provided an extra column for marking unclear cases and suggesting new category candidates. Step 2 ended with another workshop in which inter-reviewer inconsistencies were resolved and the final category model was settled. However, the category model from step 1 turned out tro be rather stable and the second step did not introduce major revisions: Changes did not affect the overall struc-

ture of the model and merely pertained to instances of some categories (e.g., the *content* category was specified more precisely (possible instance values *review*, *tutorial*, *survey* and *manual*).

3 Results – Emerging Tag Category Models

From our study data three different models emerged. Figure 1 gives an overview of the overall structure of the category model. We studied linguistic features, the relation between tags and the text of the tagged resources, as well as functional and semantic aspects of social tags.

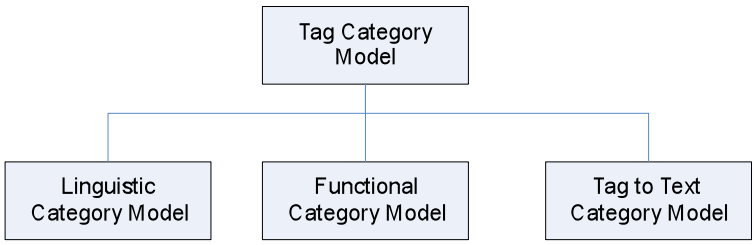

Figure 1 – Category model – overview

The following sections explain the different sub-models in more detail and report the results from our classification efforts.

3.1 Linguistic Category Model (LTCM)

In this category model we focus on linguistic aspects of tag (morpho-)syntax, orthography and lexicon. Tags are either categorised as single word tags or as multi word tags which consist of several words or phrases. Single word tags were classified according to their word class. Additionally, variations in spelling as well as neologisms and the language of the tags were noted.

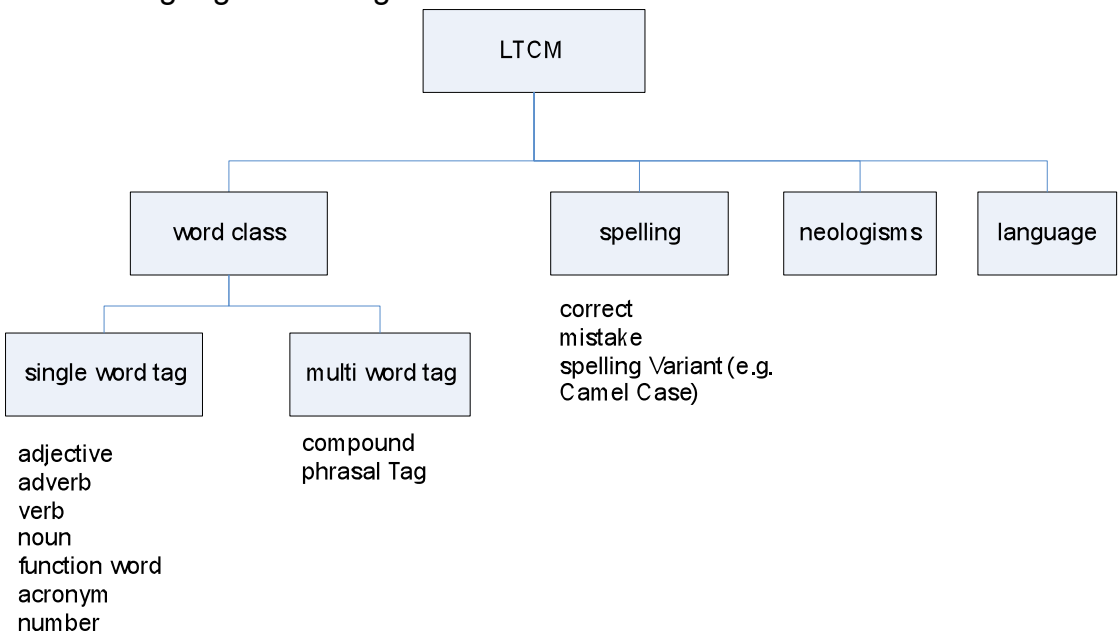

Figure 2 – Linguistic category model

3.1.1 Word Class

The majority of tags in our dataset (1191 tags) are single word tags (844 tags). However more than one quarter of all tags (347) consists of more than one word (for exact distributions see Table 1).

Number of words per tag	Occurrences	Percent total
1	844	70,87 %
2	289	24,27 %
3	46	3,87 %
4	7	0,59 %
5	2	0,17 %
6	1	0,08 %
7	0	0
8	2	0,17 %
Overall	**1191**	**100 %**

Table 1 – Distribution of word numbers per tag

The distribution of single word tags into the major word classes is summarized in Figure 3. For word class categorisation we used a fairly traditional model as the application of more detailed POS annotation tagsets (like the CLAWS4 tagset (Garside & Smith 1997) or the *Penn Treebank Tagset* (Marcus, Santorini & Marcinkiewicz 1993) seemed not appropriate for the limited breadth of linguistic phenomena under inspection. Word class ambiguity in English is a well-researched phenomenon, and it is clear that for English words with little context (as in the case of tags) this ambiguity is especially high (Jurafsky & Martin 2000, 312ff). For practical reasons, annotators decided with a bias for nouns which means that in doubtful cases the noun class was preferred. This appears to be acceptable as tags content descriptions tend to use nouns primarily.

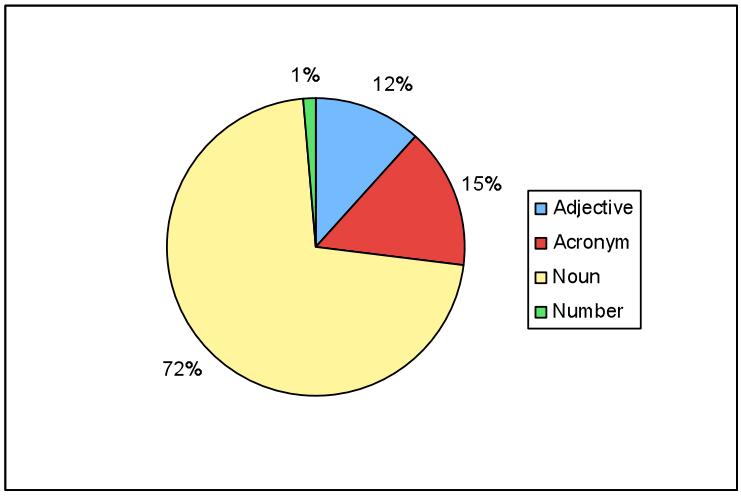

Figure 3 – Single word tags and their major word classes

Regular nouns occur most frequently (72%), followed by acronyms (15%) and adjectives (12%). Numbers[ii] are used in only 1% of all cases. They appear either in form of references to years (e.g. *1978, 2005*) or arbitrary[iii] references like *958*. Verbs, function words and adverbs are not reported in the diagram since their respective num-

bers of occurrence are too low to be represented in percent figures (0, 2 and 1 respectively).

3.1.2 Neologisms and Spelling

With the exception of 7 Spanish and 11 Italian tags, all tags are in English. Of the ten tags which are marked by the experts as neologisms, only one example withstands closer inspection: *imagingvis*. The other cases are infrequent terms which have not been invented by the taggers. Consequently, users of our system appear to be rather conservative in terms of word usage.

This conservative tagging behaviour is possibly influenced by the way the system displays previously assigned tags: According to Sen et al. (2006) main influence factors on tagging behaviour are *personal tendency* and *community influence*. Personal tendency covers factors like previous experience, knowledge and interests. The notion of *community influence* is based on the theory of social proof which states that people act the way they observe others acting (cf. Cialdini (2001)). Consequently "correct" tagging behaviour is influenced by the way the system's user interface displays tags previously assigned by other users. We also found evidence for another influence of system design on tagging practices: *CiteULike*, another social software platform for tagging scientific literature, does not allow for multi word tags and users have to adopt alternative strategies to assign a multi-word tag to a resource (cf. Kipp 2006). *Connotea* on the other hand allows separation of words within tags by spaces. Consequently hardly any *CamelCaseTags*[iv] could be found in our *Connotea* dataset.

Few spelling errors were found. Only 19 tags contain obvious spelling errors. The low error rate can possibly be ascribed to the system's tag completion algorithm, which presents possible suggestions created from previously assigned tags while users are typing.

3.2 Functional Category Model (FTCM)

The functional category model makes a distinction between subject-related tags and non-subject-related tags: Subject related tags can either describe the resource itself by giving information about author, document source or publishing date, or describe the content of the resource.

A first step towards systemizing social tags has been taken by Kipp (2007) who examines the use of non subject related tags. These tags do not show any direct relation to the text but are influenced by the user's current projects, activities and emotional state. We are following and extending Kipp's category proposal for non-subject related tags by grouping them into "affective" and "time and task related tags". Affective tags show emotional reactions to the tagged resources and can be either positive (e.g. *cool*, *fun*) or negative (e.g. *boring, dull*). A third category *tag avoidance* was added to represent tags that do not have any function, but were merely assigned because users are required to do so (e.g. *no-tag*). Fig. 4 gives an overview of our functional category model.

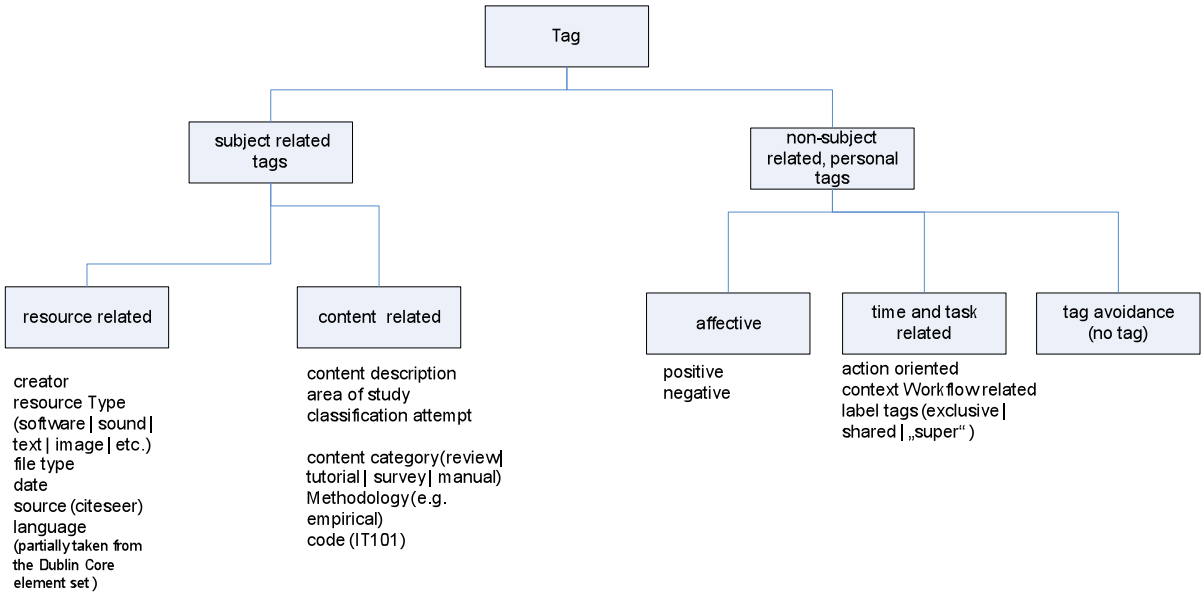

Figure 4 – Functional category model

92% of all tags can be categorized as subject related, while only 8% of all tags in our dataset were classified as non subject related (see Table 2 for rounded percent values of tag distribution in the functional model).

subject related tags	**92%**
resource related	2%
content related	98%
non-subject related tags	**8%**
affective	1%
time and task related	20%
tag avoidance	79%

Table 2 Tag distribution (rounded) – functional model

3.2.1 Subject Related Tags

The largest category in the functional model, *subject related tags* (1096 tags), is split up into 2% of *resource related tags* and 98% *content related tags*. From the resource related tags in our dataset only one tag referred to the creator of the resource, whereas 28 tags referred to a date (e.g. *2005*), to the source of the document (e. g. *citeulike, CiteSeer*) or both (e. g. *iuk2006*).

A more precise categorization of content-related tasks turned out to be a difficult task: While some tags (e.g. *resource related tags*) could be clearly allocated to a specific class, we are faced with rather difficult decisions when attempting to decide whether a tag is a simple representation of the content or an attempt to assign the resource itself to a category (i. e. a users's *classifcation attempt*). An article on collaborative filtering for example is likely to be tagged with "collaborative filtering". On a basic level this tag merely describes what the article is about, but the user could also go

one step further and file the resource under a mental folder called “collaborative filtering”.

While this does not make any difference for keyword-based information retrieval, the cognitive process behind this is rather different one: Mere content description is a simple activation of concepts in our mind that tries to capture the “aboutness” of the resource with no attempts for classification. It can even be done by simply copying and pasting what is in the title or abstract of the article. The second case reflects further elaboration strategies which place the document into a specific class of literature. Here the users actively settle on a class which is suitable to represent the tagged resource (for a cognitive analysis of tagging see Sinha (2005)). These two understandings of tagging have briefly been discussed in Coates (2005). Without futher questioning of the users directly it is hard to determine to which of these categories a tag can be assigned to. Further examples of such multiple tag interpretations from our data include: *architecture, framework analysis, prototyping, psychology, software, text mining, usability, viral marketing*.

Therefore, although initially intended, we don’t distinguish between the tag categories *Content Description, Area of Study* and *Classification Attempt*. Instead, we use the term *General Content Description* and do not make any further distinction. Figure 5 summarizes the distribution of the content related tags. Examples for other content related tag classes can be found in Table 3.

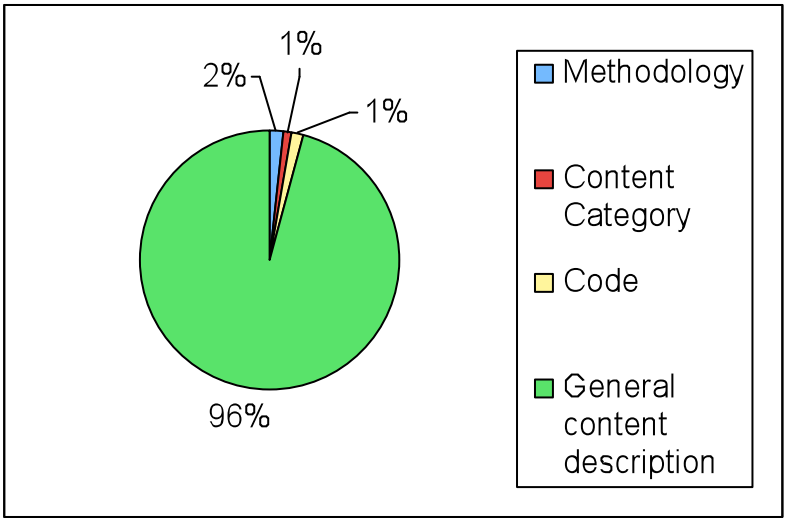

Figure 5 – Content related tags

Methodology	*theory, evaluation, argumentation, formalism, comparison, research, qualitative study, evaluation*
Content Category	*thesis, survey, review, overview, conference proceeding, tutorial, introduction, journal article*
Code[v]	*hb1, cs631, mbb806, cs431, 958, lsc895*

Table 3 – Classes of content related tags

3.2.2 Non-subject Related Tags

Time and task related tags account for 20% of the non-subject related tags of our dataset. We distinguish between *action oriented tags*, which imply some kind of action towards the resource (e.g. *readme, read, toread*), *context and workflow related tags* (e.g. *endnote, not used, uploaded-ACM17012007, _cited-by-nips2007, printed, ea_16_05_2006*) and *label tags* (see below). We discovered 13 time and task related tags and one affective tag (*ok*).

The vast majority of non-subject related tags (79%) are mere tag avoidance strategies employed by the users to circumvent the system's requirement to assign tags to resources they wish to add to the *Connotea* database. Tag avoidance can be an explicit decision of the user or a result of the workflow within *Connotea*: Examples for conscious decisions are tags like *no-tag, testtag, test* or *test1*. An alternative method to add bookmarks to the system is to import records from an existing *BibTex* or *Endnote* database. In this case the tag *uploaded* is created automatically by the system. We include this tag in the tag avoidance category, since users did not override this default tag. *Uploaded* was found 42 times (ca. 4% of all tags).

In contrast to Kipp and Campbell (2006) who studied tagging strategies in *del.icio.us* bookmarks only few "time and task related tags" like *toread* and *cool* are included in our dataset. While Kipp and Campbell claim that 16% of all tags are time and task related our dataset included just 24 or 1.3% tags of this category. Under the assumption that scientists use a special kind of language register when annotating bibliographies, this low number can possibly be ascribed to fundamental differences between scientific and standard language use (cf. Wüster 1991). Another possible explanation might be that the typical *Del.ici.ous* user has different characteristics and interests than a *Connotea* user, as the latter system clearly addresses a professional academic user community (*Connotea* being advertised as a "free online reference management for all researchers, clinicians and scientists").

3.2.3 Tags as Labels

Among the context and workflow related tasks a category *label tag* appears to emerge. We refer to a tag as an *exclusive label* when it is used rather heavily by a user but not by any other users in the system at all. Examples of *exclusive label* tags are given in Table 4.

User	**Tag**	**Number of times used in personal collection**
Linguini	958	19
fsyu2005	Timetabling	6
Mthomure	latent-semantic-analysis	7
Mthomure	image-search	12
mreddington	HFSP-funded	87
Radico	Trs	4
Wyng	Sensornet	18
Hobohm	hb1	4
Tiago	DSLs	12
Chechetka	_cited-by-nips2007	13

Mvoong	Imagingvis	8

Table 4 – Overview of exclusively used label tags

Functionally, label tags cannot be clearly assigned to a single category: *image-search* for example could be regarded as classification attempt or content description, whereas *958* remains a rather arbitrary reference.

Apart from the *exclusive label* tag we also found evidence for a label which is also used rather frequently in a personal collection of **several** users (for examples of this *shared label* type, cf. Table 5): *cs431,* for example, seems to refer to a university course in computer science which is attended by several registered *Connotea* users.

User	Tag	Number of times used in personal collection
ray178	cs431	2
vincen-trouilly	DOE	15
sailu	GO	10
greynolds	ACI	4
nsshami	cs631	4
hotzen-plotz12	cs631	19

Table 5 – Examples of shared label tags in Connotea

A third, extended function of the label was applied by one user, who used punctuation marks to create a personal classification scheme, thus organising the information in a hierarchical manner: *data::gene perturbation*, *data::sequence*, *method::transitive reduction*. However, as for the categorization of subject related tags, determining whether a tag is used as a label, a general context or workflow related tag, or a mere content description remains a tough challenge without further knowledge of the users and their respective tasks. That only a single user introduces an explicit concept hierarchy in hin/her tags could be an indicator for the hypothesis that ad hoc-usage of complex concept structures is beyond typical taggiong practices. Research on the deficiencies of the user-induced category system in Wikipedia (cf. Hammwöhner 2007) strengthens this argument.

3.3 Tag to Text Category Model (T2TCM)

This section of the paper addresses the question whether social tags tend to be taken from the full text of the tagged resource or whether users tend come up with new terms with no direct relationship with the tagged document. The idea behind this question is inspired by an information retrieval point of view: If the majority of the tags is simply taken directly from the text without any modifications, then the benefits for a full text indexing information retrieval system are marginal, as no additional metadata is created and no additional vocabulary for search queries develops[vi].

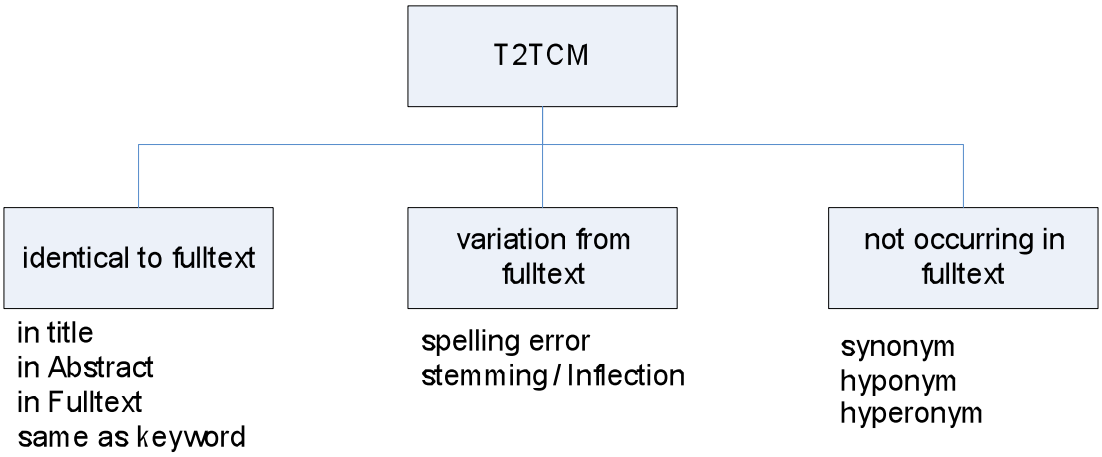

Figure 6 – Tag to Text Category Model

We classify the relationship of the tags to the full text as follows:

- *Identical* to full text – Tags either directly appear in the title, abstract, full text or as a keyword
- *Not occurring* in full text – Tags do not occur in the original text at all. Tags may be interpreted using semantic relations like synonymy, hyperonymy, hyponymy, or no obvious relation at all.
- *Variation* from full text – This occurs either in the form of a spelling error or as morphological variation of the word form (e.g. classifiers ➔ classifier).

Fig. 7 summarizes our finding for tag-to-text relationships:

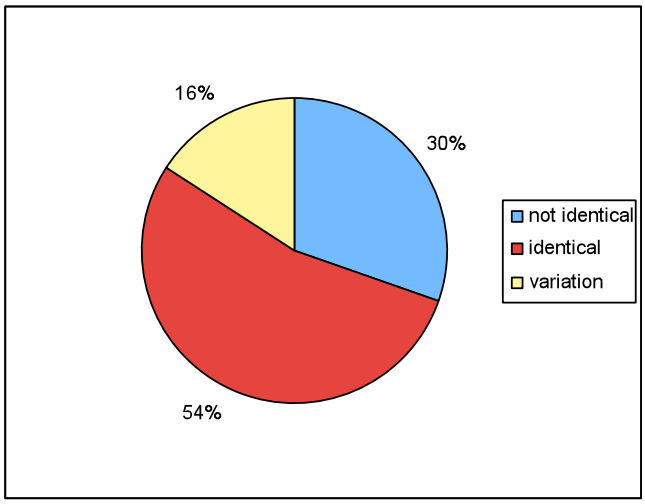

Figure 7 – Relation of tags to full text

3.3.1 Tags not Identical to Full Text

Over 30% of all tags show no relation to the text of the tagged resource at all, i.e. these tags provide some kind of novel information which cannot be provided by full text analysis of the respective documents.

3.3.2 Tags identical to full text

54% of the tags can be found in the text without any variation. The position of tag occurrence was broken down into three different categories (cf. Figure 8):

(1) Tags matching a word in the title of the resource (49%),
(2) tags matching a word in the abstract of the resource (9%), and
(3) tags matching some other word in the full text (42%)[vii].

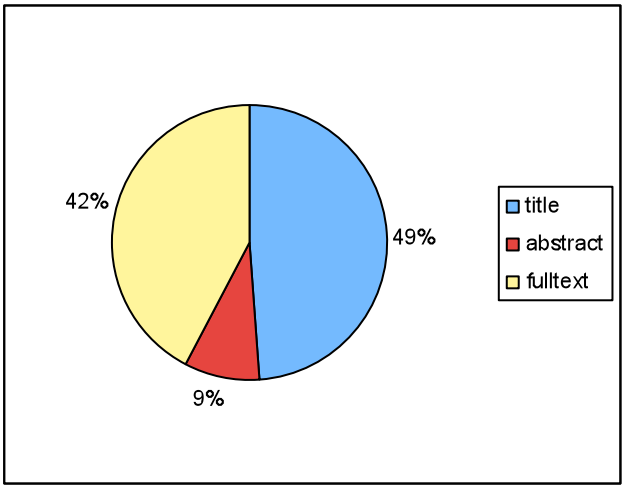

Figure 8 – Position of tag in resource

Almost half of the identical tags occur in the title of the resource. When taking into account that the title contains only a fraction of the words of a document this number seems to be surprisingly high. At the same time, *title*, *author keywords* and *abstract* are prominent features of documents that, especially in case of scientific articles, can usually be accessed without obtaining a licence for a digital library or an online journal. One possible explanation is that among those cases where tags are taken from the document title are at least some in which the tagger did not have access to the full text of the document (or did not take care to read the article even if full text was available).

3.3.3 Variation from Full Text

The tendency to rely upon the title as tag resource continues in the investigation of the tags that vary from the full text: Most tags are a variation of the title of the tagged document (63%). 14% are a variation of an existing keyword, while 10% of the tags are a variation of a word in the abstract. The remaining 13% are taken from the remaining full text of the document. We discovered three different types of variation:

(1) Spelling errors,
(2) variation of case and
(3) change of inflection (e.g. *cases* ➔ *case*).

While spelling errors are quite rare (only 2%), variation of case (46%) and change of inflection (52%) occurs more often.

3.4 Tag Category Models and Author Keywords

This section explores some of the differences between social tags and author keywords (Gil-Leiva & Alonso-Arroyo (2007), Hartley & Kostoff 2003): In addition to the tags we also classified the author keywords contained in the documents of our sample using our category models.

For those documents for which content-related tags as well as author keywords were available (see. ch. 3.2 above) we compared tags and keywords. Author keywords contain more words per keyword: While tags averaged on 1.3 words per tag, author keywords averaged on 1.8 words per keyword. At the same time, the number of tags per document was only slightly higher than 2, while on average each document contained almost 6 authors' keywords (which may be due to formal requirements imposed on authors[viii]). Thus, the maximum overlap ratio is in almost all cases bounded by the lesser number of tags when compared with the number of keywords. Looking at identical or near identical tags and keywords we found an overlap ratio of 60% relative to the minimum of the number of tags and keywords per document. The coverage of tags with respect to all authors' keywords reaches only 30% which means that almost two thirds of author keywords are not reflected in (user) tag contents.

Comparing tag and keyword contents we could observe typical thesaurus relations (cf. Kipp 2006 who employs a similar evaluation method) like broader, narrower or related terms. While both, generalisation (e.g. "RNA" (tag) versus "RNA secondary structures" (keyword") as well as specialisation (e. g. "information visualization" (tag) versus "visualization" (keyword)) can be observed, in most cases tags tend to be more general which is to be expected as tags are shorter (less multiword terms). Additionally, in some cases taggers tend to use faceted tags where authors employ (more specific) multi-word terms. Additional modifications concerning orthography or number occurred as well with a singular (tag) – plural (keyword) opposition as the most notable example (e.g. "wavelet" (tag) versus "wavelets" (keywords).

In general, taggers tend to introduce less and simpler concepts avoiding very specific terms. Although we do not have explicit evidence for this, an explanation might be that authors try to be as specific about the contents of their paper as possible (differentiation strategy with respect to a possibly huge amount of literature in the same field), while taggers try to classify the documents read by them with respect to more general categories.

3.5 Signs of Tagging Specific Language

In this section we attempt to determine whether signs of a language variety specific to social tagging can be discovered (see Crystal 2006 (esp. ch. 8/9, p. 257ff) for a current overview of "internet linguistics" and emerging varieties of internet language): The manifold means of communication that are based on digital media have led to the development of new language varieties which reflect the respective technological means and their affordances (Lee 2007) as e.g. SMS talk, internet slang, chat or

email language etc. (Abel 2000; Crystal 2006). Among the typical characteristics of these language varieties are short words as well as colloquial or dialectical expressions. Furthermore the user language constantly adapts to the quickly developing technological environment which is reflected at the lexical level (large number of neologisms). Additionally, language usage in the internet is characterized by compensation strategies for the restrictions of written communication, e.g. in form of emoticons; at the same time phenomena of spoken language are found in chat bulletin board dialogues ("written orality"). Internet users show the tendency to put less focus on punctuation and spelling rules (violation of punctuation and capitalization, typing mistakes, etc) which may be due to either the easy production circumstances of text in digital communication or to a more liberal interpretation of language rules by users of digital media. Furthermore, examples of word formations can be observed that are not typical for the English language or typical language usage, e.g. compounding (Crystal 2006, Storrer 2000, Schmitz 1994, Weingarten 1997). The following table summarizes typical characteristics of digital media language and relates them to actual findings from our dataset.

Category	**Type**	**Example**	**Occurrence**
Orthography	Violation of punctuation	data::gene expression	70x
	Violation of capitalization	DECISION TREES Controlled Theoretical Natural	95x
	Typing mistakes	Sytax	19x
Syntax	Violation of Grammar	-	-
	Ellipses	linking electronically searchable document surrogates	12x
	Assimilation	-	-
Morphology	Derivation (prefixation, suffixation and conversion)	-	-
	Back formation	-	-
	Compounding	Tagsrequired	10x
	Blend	-	-
	Sound alike slang	-	-
	Acronyms and abbreviation	TC DrmNo	121x
	Compensation strategies for non- and para-verbal communication (emoticons)	ok	1

Table 6: Typical characteristics of internet language (cf. Storrer 2000, Merchant 2001, Peele 2005)

To sum it up, concerning orthography, many of the characteristics of technology-based language variations appear: The violation of capitalization occurs very often (95 times in total). In one clear case it seems to be used as a means of highlighting terms (*DECISION TREES*). In other cases, the taggers' intention is less clear (e.g.

Controlled, Theoretical, Natural). The violation of punctuation is – at least in some cases – adopted for creating personal classification schemes (see ch. 3.2 above). Furthermore it is used as a means of compounding words. It could be the case that not all users are aware of the possibility to use multi-tag words with spaces. On the other hand, spelling errors and grammatical violation are rare – as tags do not constitute complex syntactic entities (sentences, text), this comes as no surprise. For the small amount of occurring spelling errors, a possible explanation is the system's tag suggestion algorithm. The morpho-linguistic analysis shows signs of internet language characteristics as well: The most obvious peculiarity is the prevalent occurrence of abbreviations and acronyms. Interestingly, compounding is rarely applied, as *Connotea* allows for spaces within words forming a tag (multi-tag-words). Compensation strategies for non- and para-verbal communication appear in form of so-called affective tags rather than emoticons. In our case, they are not very common. This may be due to the communicative setting: Communication within *Connotea* is rather indirect and not dialogue-oriented as email or chat communication and subject indexing of scientific literature does not appear to be highly emotional.

4 Conclusion and Further Research

In conclusion, we were able to establish a category model for tags in a scientific bibliography management scenario. This model covers linguistic features, the relation between tags and the text of the tagged resources, as well as functional and semantic aspects of social tags.

The "typical tag" is a single-word noun, taken from the title of the respective article (identical or variation), thus directly related to the respective subject. In contrast to previous studies the number of non-subject related tags remains rather low in the scientific data we observed and the full potential of tagging systems to describe qualities or aspects of resources does not seem to be used. But the absence of tags like *cool, interesting, to_read* does not mean that users who tagged the resource do not think it is cool, of interest or worthy of reading, but simply that the users did not express their ideas they may have or may not have about the resource. A possible way to elicit these ideas from users could be the addition of rating scales that measure the *interestingness* or *readworthiness* of a resource on a point scale. *CiteULike.org* is making a step in that direction by capturing the personal priority of users to read an article in their library. Influences on system usability remain an open question.

Compared to author keywords, social tags tend to introduce less and simpler concepts: Altogether, only one third of the social tags matched with (the far more numerous) authors' keywords. Moreover, tags tend to be more general and users tag their articles more general and with less words than authors.[ix]

Due to our setting, i.e. observing real data without getting hold of the users, we were not able to distinguish between tags that describe content and more elaborate tags that are intended to describe a suitable content class for the resource. Studying what

“taggers” do, how they tag content, what conscious decisions they make when they tag remains an interesting area for further investigation. This requires a controlled design of experiment where participants can be questioned about their tagging decisions.

One important outcome of the study is the observation that almost half of the tags (46%) are not found in the document text. This shows that users’ tags considerably add to the lexical space of the tagged resource. Actual retrieval effectiveness studies for tagging platforms are still missing, though.

Due to the labor intensive process of manual categorisation our dataset remains rather small. Consequently, the results need to be confirmed on a larger basis with the use of more automated techniques for analysis. Additionally, it shows that the respective system environment, e.g. tag suggestions, has a major influence on the tagging behaviour in terms of spelling errors, tag usage and creation of a specific tagging language. This extends the number of the main influential factors on tagging behaviour being *personal tendency* and *community influence* through the additional component *system influence.* Consequently, these three components should be considered in further studies.

Another area for further research is a comparative study of different tagging systems. The influence of system-related effects on tagging behaviour could be studied and related to our model. Apart from the influence of technical system charscteristics the type of tagged content plays a major role: Photos on *flickr.com* will probably be tagged using different strategies than the scientific papers which have been examined in this study.

5 Acknowledgements

Rainer Hammwöhner (Information Science, University of Regensburg) took part as an expert in the initial model-finding phase of the study and we would like to express our gratitude for his willingness to contribute his expertise to this study. We would like to thank cand. phil. Manuel Burghardt who did a great job of loading und preparing the raw data from Connotea prior to our analysis. We also appreciate the helpful comments of the anonymous reviewers on an extended abstract of this paper prior to its presentation at the 6th European Networked Knowledge Organization Systems (NKOS) Workshop, held at the 11th ECDL Conference, Budapest, Hungary.

6 References

Abel, J. (2000) *Cyberslang: Die Sprache des Internet von A bis Z* (München: C.H. Beck)

Aurnhammer, M. (2006) “Integrating Collaborative Tagging and Emergent Semantics for Image Retrieval”. Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., & Shachak, A. (2006) “Structured vs. unstructured tagging – A case study”. Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Begelman, G., Keller, P., & Smadja, F. (2006) "Automated Tag Clustering: Improving search and exploration in the tag space". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Cialdini, R. B. (2001) *Influence Science and Practice* (Boston, MA: Allyn and Bacon)

Coates, T. (2005) "Two cultures of fauxonomies collide…" Available online: http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide/ . Last access: May 8, 2008.

Crystal, D. (2006) *Language and the internet.* 2nd edition. (Cambridge: Cambridge University Press)

Damianos, L., Griffith, J., & Cuomo, D. (2006) "Onomi: Social Bookmarking on a Corporate Intranet". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Dennis, B. (2006) "Foragr: Collaboratively Tagged Photographs and Social Information Visualization". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Farrell, S., Lau, T. (2006) "Fringe Contacts: People-Tagging for the Enterprise". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Foskett, A. C. (1997) *The subject approach to information* (5th ed., repr.) (London: Library Association Publishing)

Feinberg, M. (2006) "An Examination of Authority in Social Classification Systems". Paper presented at the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Austin/TX, November 2006 Available online: http://dlist.sir.arizona.edu/1783/ . Last access: May 8, 2008.

Furnas, G.W., Landauer, T. K., Gomez, L. M., & Dumais S. T. (1987) "The vocabulary problem in human-system communication". *Commun. ACM.,* Vol. 30, 964-71

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora (pp. 102-121). London: Longman.

Gil-Leiva, I., & Alonso-Arroyo, A. (2007) "Keywords given by authors of scientific articles in database descriptors". *J. Am. Soc. Inf. Sci. Technol.,* Vol. 58, 1175-87

Golder, S., & Huberman, B. A. (2006) "The Structure of Collaborative Tagging Systems". *Journal of Information Science,* Vol. 32, 198-208

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005) "Social Bookmarking Tools (I): A General Review". *D-Lib Magazine,* Vol. 11, Num. 4, Available online: http://www.dlib.org/dlib/april05/hammond/04hammond.html . Last access: May 8, 2008.

Hammwöhner, R. (2007). "Interlingual Aspects Of Wikipedia's Quality." Paper presented at the 12th International Conference on Information Qualiy, ICIQ 2007. Available online: http://mitiq.mit.edu/iciq/ICIQ/iqdownload.aspx?ICIQYear=2007 . Last access: May 8, 2008.

Hartley, J. & Kostoff, R.N. (2003) "How useful are `key words` in scientific journals?". *Journal of Information Science,* Vol. 29, 433-38

John, A., & Seligman, D. (2006) "Collaborative Tagging and Expertise in the Enterprise". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Jurafsky, D., & Martin, J. H. (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. (San Francisco: Prentice Hall)

Kipp, M. (2006) "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords". *Canadian Journal of Information and Library Science*. Available online: http://dlist.sir.arizona.edu/1533/. Last access: 14 September 2007.

Kipp, M. (2007) "@toread and cool: Tagging for time, task and emotion". Paper presented at the 8th Information Architecture Summit, Las Vegas. Available online: http://eprints.rclis.org/archive/00011414/ . Last access: May 8, 2008.

Kipp, M. E. I., & Campbell, D. G. (2006) "Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices". Paper presented at the 2006 Annual Meeting of the American Society for Information Science and Technology, Austin.

Kroski, E. (2005) "The hive mind: Folksonomies and user-based tagging" Available online: http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/ . Last access: May 8, 2008.

Lakoff, G. (2005) *Women, Fire and Dangerous Things* (Chicago: University of Chicago Press)

Lee, C. K.-M. (2007) "Affordances and Text-Making Practices in Online Instant Messaging". *Written Communication, 24(3)*, 223-249.

Lancaster, W. (1993) *Information retrieval today* (Arlington: Information Resources Press)

Merchant, Guy (2001) "Teenagers in cyberspace: an investigation of language use and language change in internet chatrooms". *Journal of Research in Reading*, Vol. 24, Num. 3, 293-306

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993) "Building a Large Annotated Corpus of English: The Penn Treebank". *Computational Linguistics*, 19(2), 313--330.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006) "HT06, tagging paper, taxonomy, Flickr, academic article, to read" In *Proceedings of the seventeenth conference on Hypertext and hypermedia*. (New York: ACM Press), pp. 31-40

Peele, A. (2005) "The Prevalence of the English Language in Communicating on the Internet" *Revista de Informatică Socială,* Vol. 2, Num. 3, 82-7. Available online: http://www.ris.uvt.ro/Publications/Iunie%202005/Pele.pdf . Last access: May 8, 2008.

Schmitz, P. (2006)" Inducing Ontology from Flickr Tags". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Schmitz, U. (1994) "Neue Medien und Gegenwartssprache: Lagebericht und Problemskizze" *Osnabrücker Beiträge zur Sprachtheorie (OBST)* 50 (1995), 7-51. Available online: http://www.linse.uni-due.de/linse/publikationen/n_medien_gegenwartsspr.html . Last access: May 8, 2008.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., & Osterhouse, J. (2006) "Tagging, communities, vocabulary, evolution". Paper presented at CSCW 2006, Banff.

Shirky, C. (2005a) "Folksonomies + controlled vocabularies" Availabe online: http://many.corante.com/archives/2005/01/07/folksonomies_controlled_vocabularies.php . Last access: May 8, 2008.

Shirky, C. (2005b) "Ontology is overrated: categories, links and tags" Available online: http://shirky.com/writings/ontology_overrated.html . Last access: May 8, 2008.

Sinha, R. (2005) "A cognitive analysis of tagging" Available online: http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html . Last access: May 8, 2008.

Trant, J., Wyman, B. (2006) "Investigating social tagging and folksonomy in art museums with steve.museum". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Storrer, A. (2000) "Schriftverkehr auf der Datenautobahn. Besonderheiten der schriftlichen Kommunikation im Internet". In *Neue Medien im Alltag. Begriffsbestimmungen eines interdisziplinären Forschungsfeldes,* edited by Gerd Günter Voß (Leverkusen: Leske + Budrich Verlag)

Voß, J. (2007) "Tagging, folksonomy & co – Renaissance of manual indexing?". Paper presented at Open Innovation – neue Perspektiven im Kontext von Information und Wissen, 10th International Symposium for Information Science, Cologne.

Weingarten, R. (ed.) (1997) *Sprachwandel durch Computer* (Opladen: Westdeutscher Verlag)

Wüster, E. (1991) *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* (Bonn: Romanistischer Verlag)

Xu, Z., Fu, Y., Mao, J. & Su, D. (2006) "Automated Tag Clustering: Improving search and exploration in the tag space". Paper presented at WWW2006, Collaborative Web Tagging Workshop, Edinburgh.

Yew, J., Gibson, F., & Teasley, S. (2006) "Learning by tagging: group knowledge formation in a self-organizing learning community" In *Proceedings of the 7th international conference on Learning sciences,* (Bloomington: International Society of the Learning Sciences), pp. 1010-1

---

[i] The sample for step one included 706 articles and 2426 tags.

[ii] A tag was counted as a number, when the tag did not contain any letters at all, so the cases like *p2p* and *3D* were not considered.

[iii] Arbitrary in the sense that the annotator cannot interpret the correct meaning with information available for tag classification.

[iv] CamelCase, or medial capitals denotes a practice of forming compound words where initial capitals are retained in the new compound word, e.g. RealPlayer or DaimlerChrysler. Camel case is a common practice in programming.

[v] In most cases, codes appear to refer to academic course codes ("cs101").

[vi] Although there still *is* more information in a tag than in a word in the full text, simply due to the fact that the tagger has chosen the word to be worthy of describing the resource.

[vii] The analysis is carried out hierarchically, i.e. if a tag is found in the title, then abstract and full text are no longer searched (and so on).

[viii] For a discussion of keywords in scientific articles, see Hartley, J. & R.N. Kostoff (2003) and Gil-Leiva & Alonso-Arroyo (2007).

[ix] It would be desirable to take *experts'* keywords (as found in bibliographic or library information systems) as a third way of intellectually annotating articles into account, e.g. by looking up keywords and classification codes for ICT-related articles in databases like INSPEC or Compuscience.