Verbatim and semantic imitation in indexing resources on the Web:

A fuzzy-trace account of social tagging

Paul Seitlinger

Graz University of Technology, Graz, Austria

Tobias Ley

Tallinn University, Tallinn, Estonia

Dietrich Albert

Graz University of Technology, Graz, Austria

University of Graz, Graz, Austria

Published as: Seitlinger, P., Ley, T., & Albert, D. (2015). Verbatim and semantic imitation in indexing resources on the Web: A fuzzy-trace account of social tagging. *Applied Cognitive Psychology*, *29*(1), 32–48. doi:10.1002/acp.3067

Acknowledgments: This work has been partially funded by the Austrian Science Fund (FWF): P 25593-G22, and the EU FP7 under the Learning Layers project (http://www.learning-layers.eu).

Abstract

Social Tagging is a widespread phenomenon on the Web allowing users to tag resources, such as photos, by freely chosen labels. Imitation of other users' tagging behavior is deemed to increase the inter-individual consistency of tag assignments. Both verbatim and semantic mechanisms have been proposed where the first case suggests reuse of the exact words, and the second reuse of the concept without necessarily the word form. Here, we present a multinomial model of assigning tags that integrates these perspectives. Based on the fuzzy-trace theory, the model includes separate parameters for the retrieval of verbatim and gist traces of encountered tags. Results of two experiments demonstrate that both types of memory traces contribute substantially and independently to tag productions. We conclude that the imitation of tags may be explained within one fuzzy-trace framework that contributes to our understanding of emergent phenomena in social information systems.

Keywords: Memory, Imitation, Fuzzy-trace theory, Multinomial modeling, Social Web

1 Introduction

In the last ten years, there has been a renaissance of manual resource indexing on the Web in form of a specific variant called Social Tagging (Voss, 2007). Social Tagging (ST) is a widespread functionality in social Internet environments providing users the opportunity of freely assigning labels (so called tags) to collaboratively collected Web resources, such as books (http://librarything.com), photos (http://flickr.com) or sound files (https://soundcloud.com). Similar to document indexing in a library, the purpose of ST is to structure resources and to aid the retrieval of relevant content on the Web. While in the context of libraries, the importance of providing index terms for an efficient information retrieval has long been agreed (e.g. Bates, 1998), it is unclear whether ST can play a similar role for content on the web.

The main difference between document indexing in a library and tagging on the Web is that users of an ST environment are not restricted to a controlled vocabulary but can apply any tag they find appropriate. Because of this low effort, ST has become very popular among users and therefore, promises high levels of scalability even for large corpora of content as we find on the Web. Despite this uncontrolled way of indexing, research on ST has revealed an emerging agreement on at least a few tags of a resource (e.g. Golder & Huberman, 2006). This emergence of stable patterns in social media has been attributed to the users' tendency to imitate others (Sakamoto, Ma & Nickerson, 2009), such as imitating a small subset of other users' tags (e.g. Sen et al., 2006).

To explain the tendency towards imitation and its effects, several generative models have been suggested, which computationally specify processes of tag imitation, such as preferential attachment (Fu, Kannampallil & Kang, 2009). One class of these models aims at predicting *verbatim imitation*, i.e., the probability of reusing a tag (e.g. Cattuto, Loreto & Pietronero, 2007, Dellschaft & Staab, 2008; Halpin, Robu & Shepherd, 2007). This conceptualization of imitation is only concerned with predicting reuse of the surface form of the tag, i.e., its exact wording. The second approach goes beyond the word level and models *semantic imitation*: it is not necessarily the tag that is reused, but the tagged concept (Fu et al., 2009; Fu, Kannampallil, Kang & He, 2010). The assumption is that certain tags, such as *moment* and *time*, prompt a user to think of a particular topic, which in turn leads to the production of new but semantically related tags, such as *instant*, or to the reuse of old ones, such as *moment*.

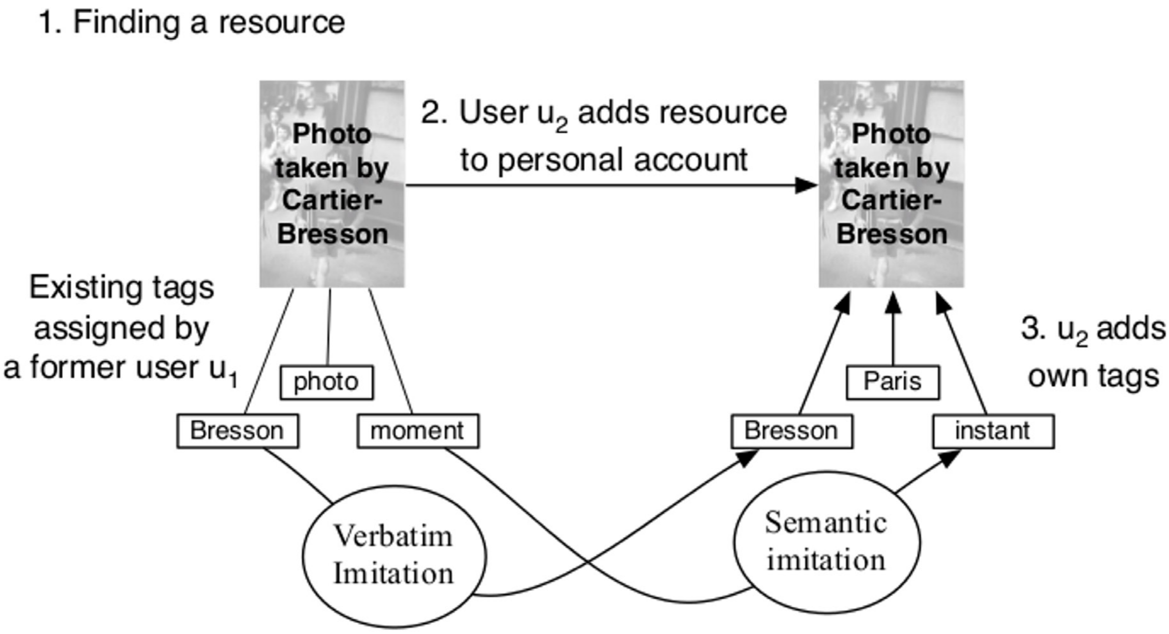

Figure 1. Example to illustrate verbatim and semantic imitation of tags

Figure 1 provides an illustrative example of the two different types of imitation in the context of a tag-based photo sharing system. In this example and in the remainder of the article, we focus on cognitive processes taking place after a user has decided describing a Web resource with tags. In this case, a user u_2 finds an interesting photo taken by the artist Henri Cartier-Bresson. A former user u_1 has added the photo to the system and has assigned the three tags *photo*, *Bresson* and *moment*, which are visible for subsequent users. u_2 decides adding the photo to her personal account and in turn annotates the photo with the three tags *Bresson*, *instant* and *Paris*. While *Bresson* is a verbatim imitation of u_1's tag assignments, *instant* is a semantic imitation only sharing semantic aspects with one of u_1's tags, i.e., *moment*. Finally, the tag *Paris* can be regarded an idiosyncratic assignment that does not appear to be caused by verbatim or semantic imitation. By assigning tags, the user builds up a personal indexing system for the pictures she finds on the Web.

Whereas current models of ST are based either on verbatim or semantic imitation, psychological research on dual-recall conceptions of memory retrieval clearly shows that humans draw on both these mechanisms when retrieving from memory. For instance, fuzzy-trace theory (FTT; e.g. Brainerd & Reyna, 2005, 2010) integrates the retrieval of episodic traces storing surface and semantic features (so-called verbatim traces) as well as the retrieval of episodic traces only storing semantic features (so-called gist traces) in one dual-recall theory. Similarly, studies explaining document indexing in a library find that indexers apply a mix of perceptual (verbatim) and conceptual (semantic) strategies to assign terms to documents (e.g. Bertrand, Cellier & Giroux, 1996; David, Giroux, Bertrand-Gastaldy & Lanteigne, 1995; Farrow, 1995).

In this light, current models of tagging take a limited view of the processes underlying tag imitations. Our goal is therefore to suggest and validate a model of tag imitations that integrates verbatim and semantic imitation within one common theoretical framework and allows estimating the contributions of each of the two forms of imitation. Such a model is important for at least two reasons. First, successful models of cognitive user behavior can be implemented in form of tag recommender or feedback mechanisms that are important services to support users' indexing behavior (Jäschke, Marinho, Hotho, Schmidt-Thieme & Stumme, 2008). Secondly, employing such model would create a remarkable opportunity for research in cognitive and linguistic science to complement findings from lab studies with findings from the Web (e.g. Glushko, Maglio, Matlock & Barsalou, 2008; Goldwater, Markman & Stilwell, 2011).

In the next section, we develop such an integrative model built upon Fuzzy-trace Theory (FTT). We then validate the model and estimate the influence of verbatim and semantic imitation in a task of tagging resources on the Web. We finally apply the model in a second more open field setting.

2 A fuzzy-trace framework for memory-based tag imitations

So far, simulation studies have either implemented verbatim models (e.g. Cattuto et al., 2007; Dellschaft & Staab, 2008) or the semantic model (Fu et al., 2009). Both have independently succeeded in replicating statistical patterns typically observed in ST-environments, such as the power law distribution of tags associated with a Web resource (e.g. Halpin et al., 2007). Moreover, results of experimental studies of Fu et al. (2010) and Fu & Dong (2012) have been well in accordance with the semantic imitation approach.

However, no effort has been made to pit these models against each other in one single study to dissociate the contributions of verbatim and semantic processes. Such an analysis is hampered by the fact that only two response categories can be observed: the category of reusing a tag t, t_r, and the category of not reusing the tag, t_n. Referring to the example of Figure 1, only the tag *Bresson* belongs to t_r and both the tag *instant*, which is a semantic imitation of the existing tag *moment*, and the tag *Paris,* an idiosyncratic assignment, belong to t_n. Table 1 shows that different latent processes may underlie the category t_r. As the third column of Table 1 (labeled “Tag reused”) reveals, the total probability of reusing a tag can be broken down into three different conditional probabilities: A tag can be reused given either verbatim or semantic imitation or no imitation has taken place. In the latter case, the assignment of a tag is not affected by other users’ tags, and we denote it as an idiosyncratic assignment. In some cases, this may also result in the accidental reuse of another user’s tag. The fourth column of Table 1, then, shows how the probability for the category t_n, not reusing a previously seen tag, can be broken down. We assume it cannot be attributed to a verbatim process, but could be caused by semantic imitation: For instance, as demonstrated by the example above, the new tag *instant* may be the consequence of a semantic imitation of the previously encountered tag *moment*. Finally, t_n can also be caused by an idiosyncratic assignment.

Insert Table 1 **about here**

The psychological method of choice to measure the probability of latent processes underlying categorical data is the technique of multinomial processing trees introduced by Riefer and Batchelder (1988; see also Batchelder & Riefer, 1999; Erdfelder et al., 2009). In this article we therefore propose and test a multinomial model of ST that includes parameters for latent constructs representing verbatim and semantic imitation as well as idiosyncratic tagging behavior. In this way, we aim at estimating the influence of verbatim and semantic imitation on the observable probabilities of $p(t_r)$ and $p(t_n)$.

2.1. A fuzzy-trace interpretation of verbatim and semantic imitation

The model we propose is directly derived from FTT, which differentiates between verbatim and gist traces of learning episodes in long-term memory by distinguishing different recall states with respect to read words (e.g. Brainerd & Reyna, 2010; Gomes, Brainerd & Stein, 2012), such as tags in a photo-sharing environment. In Figure 1, the first tag assignment of u_2, the tag *Bresson*, belongs to the category of reusing an old tag, t_r. We assume that in rare cases, this can result from an idiosyncratic tag assignment in a *no recall* state U (described below). More likely, it will be caused by a verbatim imitation in the *perfect recall* state L. The latter state is associated with a verbatim trace that stores semantic features and surface details of read tags. The verbatim trace supports a recollective, direct retrieval from memory (e.g. Kintsch, 1970) that reinstates the tag's surface during a tag assignment and allows for a verbatim imitation. In FTT, the probability of entering state L and thus, directly accessing the verbatim trace of an encountered tag (without activating, searching and comparing traces of other tags) is represented by the parameter *D*. Table 2 maps the recall states and probabilities of entering these recall states on the different types of imitation that in turn result either in t_r or t_n. The first row in the table summarizes this mapping for the *perfect rec*all state *L* underlying a verbatim imitation.

Insert Table 2 **about here**

In Figure 1, we have illustrated a semantic imitation by u_2's tag *instant* that shares semantic features with u_1's tag *moment*. Referring to FTT, this tag assignment bears on a *partial recall* state *P*, in which semantic but not surface features of an encountered tag are stored in form of a gist trace. During recall, a gist trace supports semantic processing, specifically, a non-recollective retrieval that consists in a "reconstruction operation plus a slave familiarity judgment operation" (Brainerd & Reyna, 2010, p. 428). During reconstruction, a mental search set of words (e.g. Wixted & Rohrer, 1994) is generated that share semantic features with the stored gist. For instance, if u_2 memorizes the gist of the encountered tag *moment* during a tag assignment he will be able to reconstruct a set of semantically related words, such as {*moment*, *time*, *instant*, *second*, *now*}. However, this reconstruction operation does not provide item-specific information and therefore, a subsequent familiarity judgment (e.g. Brainerd, Wright, Reyna & Payne, 2002; Brainerd & Reyna, 2010) is assumed to select one of the reconstructed words for the tag assignment. If the selected word is an actually encountered tag (e.g. *moment*), we call the familiarity judgment positive and use the symbol P_r to denote a *partial recall* state that causes semantic imitation resulting in the category t_r. If, however, the selected word has not been encountered (e.g. *instant*) we call the familiarity judgment negative and use the symbol P_n to denote a *partial recall* state that also causes semantic imitation but – this time – results in the category

t_n. The probabilities of entering P_r and P_n are given by $(1-D)RJ$ and $(1-D)R(1-J)$, respectively, where R represents the proportion of reconstructed tags and J the proportion of tags that are reconstructed and actually selected for a tag assignment. The second and third row of Table 2 summarize how the two partial recall states P_r and P_n, which both underlie semantic imitation, lead to the categories t_r and t_n, respectively.

Without encountering a particular tag t, a user is assumed to be in a *no recall* state U and the probability of recalling t during a tag assignment is close to zero: In case of tagging a photo, there is still a chance of selecting t in state U since different people can apply the same label for an object even if they name the object independently of each other. In our model we account for this chance by the parameter G that represents the probability of coincidental guessing during idiosyncratic tagging and also results in the response category t_r. Hereinafter, the symbols U_r and U_n denote *no recall* states associated with the categories t_r and t_n, respectively. As the last two rows of Table 2 show, entering the recall states U_r and U_n is given by $(1-D)(1-R)G$ and $(1-D)(1-R)(1-G)$, respectively.

Before we present our research questions about the impact of verbatim, semantic and idiosyncratic imitation on tagging photos, the next two sections describe the applied paradigm and corresponding multinomial model to measure these types of imitation and to specify our research questions by means of the parameters D, R, J and G.

3 Empirical test of the fuzzy-trace model of social tagging

3.1 The adapted STTT-procedure

To identify the four parameters D, R, J and G, we apply an adapted STTT-procedure, originally introduced by Brainerd et al. (2002). It yields a sufficient number of observable response categories to estimate the FTT parameters and to perform a reliable model test. The original STTT procedure consists of a study phase S and three consecutive test phases T_1, T_2

and T_3. In phase S, the participants study a list of words and before each of the three test phases T_i, they solve simple arithmetic problems for one minute. These one-minute distractor tasks serve as buffering activities to empty primary memory and to ensure retrieval from secondary memory. In each test phase T_i, the participants perform a free (or cued) recall and get the instruction to write down as many words as they can remember from phase S.

In the experiments reported below, phase S is embedded into a tag-based information search to realize an incidental learning condition. The participants make use of the Internet environment *http://picasa.com*, which provides short descriptions and representative photos of specific artistic photographers. Additionally, we have assigned several tags to each photo. The participants' task is to search through the whole collection and to select interesting photos.

Similar to the original STTT paradigm, before each test phase the participants perform a one-minute distractor task of solving simple multiplication problems. During each test phase, the participants are presented all photos of phase S without presenting tags and are asked to tag them either with own words or tags remembered from phase S. We instructed the participants to collect photos and describe their semantic content as well as additional thoughts coming to mind so that they themselves or others would more easily find them later. With regards to different types of motives for tagging, we assume this instruction mainly evoked the motive of personal and collaborative curation as described by Marlow, Naaman, Boyd and Davis (2006). There are certainly other motives for using social tagging systems (such as social signaling) and the quality of tags varies with different motives (e.g., Marlow et al., 2006; Nov, Naaman & Ye, 2008). However, in this study (experiments 1 and 2) we focus on curation that we deem central for the application of tagging to organize resources on the Web, similar to the use of professional indexing terms in libraries. In the general discussion we will also describe future experiments that will examine the impact of different motives on the relative contributions of verbatim and semantic imitation.

A significant challenge we faced in our design was to balance the trade-off between a theory-driven approach towards measurement on the one hand and the ecological validity on the other. This is the reason why participants had to engage in a task of re-tagging the same photo three times. While this may be rather unusual for a real-world social tagging system, our aim was to measure memory processes involved in freely assigning tags to photos in a theory-guided way. Applying models that have been validated previously allows us to interpret the results of the parameter measures in terms of the FTT constructs. As all models of free recall require several test phases, we feel the STTT procedure gave us the best trade-off between ecological validity and control to measure verbatim and semantic processing in tagging photos. We did consider more elaborate models (such as the $S_1T_1T_2$,S_2T_3,S_3T_4 procedure requiring three study phases (S_i) and four test phases (T_i); Brainerd and Reyna, 2010), but found that these would depart even more from a natural use of social tagging systems. Once the parameters have been validated with the present design, we will then seek to relax the paradigm to take more natural interactions into account.

3.2 A fuzzy-trace processing tree tailored to the adapted STTT procedure

The STTT-procedure results in relative frequencies of different response categories. A participant can reuse a particular tag of phase S in a particular test phase (represented by the symbol r) or not (represented by n). Given the three test phases T_i, this binary coding yields eight (= 2^3) possible response categories: rrr, rrn, rnr, nrr, rnn, nrn, nnr and nnn. For instance, the abbreviation rrn denotes the response category of a participant reusing a tag in the test phases T_1 and T_2 but not in T_3.

The processing tree in Figure 2 shows the multinomial model applied to estimate the parameters *D*, *R*, *J* and *G*. The model corresponds to the equations in Table 2 except the two additional parameters J_{2r} and J_{2n}. J_{2r} represents the conditional probability of selecting a reconstructed tag for a tag assignment (i.e., a positive familiarity judgment) given this process has been positive in the previous test phase (T_1 or T_2). By contrast, J_{2n} is the probability of a positive familiarity judgment given this process has been negative in the test phases T_1 or T_2. We have introduced them to account for variations in the familiarity judgment across the three test phases, resulting in the parameter space $\mathbf{\Omega}_1 = [0, 1]^6$.

Within the processing tree, the rectangle at the far left represents a photo of phase S and the column on the right side lists the different response categories (across the three test phases). In between, the circles along the paths represent different sequences of learning states leading to eight unique response categories. The indices of the learning states *P* and *U* are either r or n, which symbolize "tag has been reused" or "tag has not been reused" within the corresponding learning state. More than one path of learning states, e.g. the first (L−L−L) and second (P_r−P_r−P_r), can lead to the same response category (in this case, rrr). We have arranged the learning states in the three intermediate columns representing the three test phases T_i. The FTT parameters are attached to the lines and represent the probability of a transition from one to another learning state. By multiplying the state transitions for each path in the processing tree, we have derived eight model equations (multiplicative terms) according to Hu and Batchelder (1994). See Appendix A for an example of one equation.

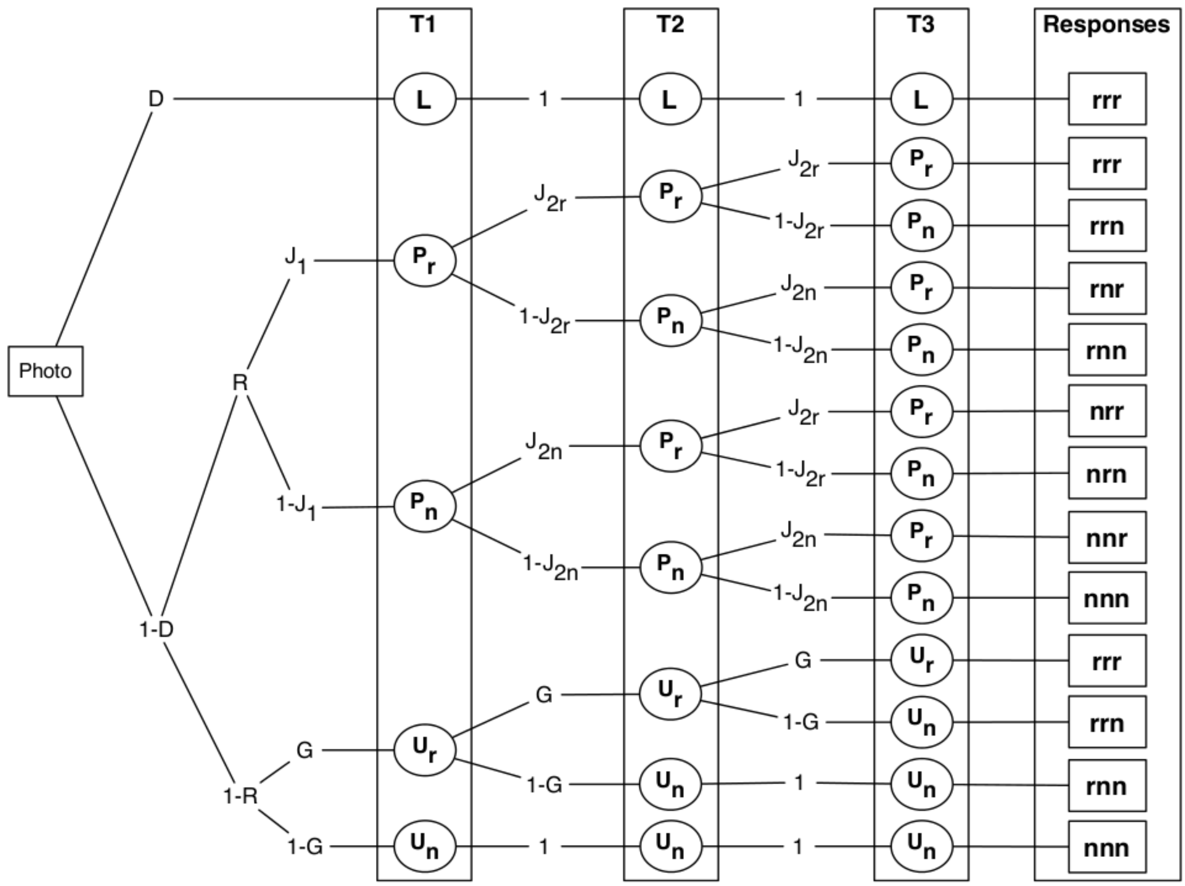

Figure 2. Multinomial processing tree to discern verbatim and semantic imitation
Note. Model parameters: *L=Perfect recall* state; P_r, P_n =*Partial recall* states that lead to t_r and t_n, respectively; U_r, U_n=*No recall* states that lead to t_r and t_n, respectively; *D*=Direct retrieval, *R*=Reconstruction, J_1=Familiarity-based judgment in the first test phase, J_{2r}, J_{2n}= Familiarity-based judgment in the first or second test phase given the judgment has been positive and negative in the previous test phase, respectively; *G*=Guessing; T1, T2, T3 represent the first, second and third test phase, respectively

The model defines a one-to-one mapping of the parameter space $\mathbf{\Omega}_1$ in the set of all possible category probabilities and hence, is identifiable (e.g. Erdfelder et al., 2009). Moreover, the number of free parameters (six) is smaller than the number of empirical probabilities (eight) allowing for a test of the model's data fit. To measure the goodness-of-fit of the model, i.e., the divergence between observed and predicted category frequencies, the power-divergence measure G^2 is used, following a χ^2-distribution asymptotically (e.g. Riefer & Batchelder, 1988).

After having introduced the multinomial model and corresponding STTT-paradigm, we are now able to raise our research questions and to phrase our predictions in terms of the model's parameters.

3.3 Research questions

In two experiments, we analyze as to whether the model we have introduced in the previous section, which combines verbatim imitation and semantic imitation, can explain the process of assigning tags. This leads to our first research question:

- Does the modified fuzzy-trace model of ST with the parameter space $\mathbf{\Omega}_1$ account for the imitation of tags? In particular, does the processing tree shown in Figure 2 fit the data gathered by means of the STTT paradigm? (RQ1)

As none of the previous models of social tagging could be used to independently estimate the amount of verbatim and semantic imitation, we then identify and estimate the parameters of Table 2 to assess the contributions of each of these to users' tag assignments. The second research questions therefore is

- To what extent are verbatim imitation (proportion of responses based on the learning state L) and semantic imitation (proportion of responses based on the learning states P_r and P_n) involved in the imitation of tags? (RQ2)

By applying the model, we then examine the impact of particular design features in an ST-environment, such as size and semantic arrangement of tags, on verbatim and semantic imitation, in order to explore opportunities of influencing users' tag choices. At the same time, and to derive further evidence for the validity of our model, we use the manipulations of the design features to explore whether the processes quantified by the parameters can be functionally dissociated. For instance, if *D* and *R* quantify probabilities of processes involved in tagging that are independent of each other, the variation of a design feature should be associated with differential effects on *D* and *R*. This leads to our third research question, which we will particularly focus on in our second experiment:

- Is it possible to influence the parameters of the parameter space $\mathbf{\Omega}_1$ by design features in ST systems? (RQ3)

Next, we present the two experiments that we have conducted to answer these research questions.

3.4 Experiment 1

In this experiment, we address the first two research questions as to whether the FTT-model of ST accounts for the imitation of tags (RQ1) and how strong the contributions of verbatim and semantic imitation are (RQ2). We assume that the proposed model in Figure 2 accounts for the tags produced by the participants in the STTT procedure. This assumption is based on a pilot study (Seitlinger & Ley, 2012) that we have conducted to check for the feasibility of the STTT procedure and to improve a preliminary dual-memory model of tagging. In the present study, we test the first hypothesis statistically by performing two sorts of tests. First, we analyze as to whether the model accommodates the empirical data and second, we compare this model fit to the fit of alternative models that assume fewer processes (e.g., only gist-based reconstruction and a subsequent familiarity judgment as well as guessing) and thus, include fewer parameters. The rationale underlying the comparison with alternative and less extensive models is as follows. By assuming two independent processes – a verbatim and a gist-based retrieval – the suggested model of Figure 2 is a dual-process model (Yonelinas, 2002). In the course of many cumulative experiments, research on FTT has gathered a wealth of data speaking in favor of both processes. However, as we have outlined in the introduction, this dual-process account of reminding and (re)producing previously presented words, has not been adopted by research on social tagging. Therefore, we compare the model depicted in Figure 2 with simpler models that either exclude the verbatim or reconstructive process to examine as to whether the dual-process assumption is actually necessary for modeling the reuse of tags. As a consequence, we compared the following models.

- *FTT Imitation Model (2SDPg):* Beyond assuming a verbatim and gist-based process, our FTT-based model decomposes the latter process into a reconstruction operation and a familiarity judgment. The additional parameter G accounts for coincidental guessing. Taken together, we denote this model a 2-step dual process plus guessing (2SDPg) model, with the six free parameters D, R, J_1, J_{2r}, J_{2n} and G.
- *Verbatim Imitation Model (1SDPg):* The first alternative model is a 1-step dual process plus guessing (1SDPg) model, supposing a verbatim process and a one-step familiarity judgment as well as guessing, resulting in the five free parameters D, J_1, J_{2r}, J_{2n} and G.
- *Semantic Imitation Model (2SOPg):* The second is a 2-step one process plus guessing (2SOPg) model, including the two-step reconstruction-familiarity process as well as guessing and excluding the verbatim process, in turn resulting in the five free parameters R, J_1, J_{2r}, J_{2n} and G.
- *Simplified FTT Model (2SDP):* Finally, we test the fit of the original FTT-model (Brainerd et al., 2002) not taking into account the possibility of guessing, denoted 2SDP model with the parameters D, R, J_1, J_{2r} and J_{2n}. The model equations of the four models are given in the appendix B (Table 10).

Our assumption is that the 2SDPg model *yields a significantly better fit than the three alternative models (hypothesis 1.1).*

Besides examining the relative contributions of verbatim and gist traces of encountered tags to the production of own tags, we used this first experiment to also assess the strengths of the memory traces in comparison to a baseline, namely memory traces left by expert-generated keywords. As mentioned in the introduction, these indexing keywords have been found to be important cues when searching for information (e.g. Bates, 1998). They are produced by expert indexers who draw on a controlled documentary language to express the main ideas of a document (Bertrand et al., 1996). To the contrary, tags are frequently generated by laymen and not part of a controlled vocabulary. Hence, users may not draw as much attention to tags as to keywords because they are aware that laymen have produced the former category of information. This assumption is in line with work of Gerjets and Kammerer (2010) who have found people searching the Web to be very sensitive to the source of information and to draw more attention to sources they deem reliable.

To further examine RQ2, we therefore introduce in this study expert-generated keywords that are displayed alongside the social tags, and we use a model for the processing of those keywords as a baseline condition. Note that we draw the tags and keywords from the same pool of words and only vary the pretended source from which the words are allegedly drawn. By comparing the processing of user-generated tags and expert-generated keywords, we analyze the extent to which the strengths of traces of the two classes of words correspond to each other. If we find only a small divergence between traces of user-generated tags and traces of expert-generated keywords, then this would substantiate the claim that social tags are in fact important information cues in information search on the Web. *Statistically, we explore whether significant differences exist in estimates of the parameters D, R, J and G between user-generated tags and expert-generated keywords (explorative hypothesis 1.2).*

3.4.1 Participants

24 psychology students at University of Graz (11 female) with an average age of 31.4 years (SD=10.1, ranging from 18 to 60 years) participated in the experiment for course credits

or for a 10€ theater coupon. Each participant was tested individually in a laboratory session lasting approximately 52 minutes.

3.4.2 Materials

For the study phase S, participants used an application providing information about six visual artists: Marina Black, Henrie Cartier-Bresson, Saul Leiter, Astrid Kruse-Jensen and Helen Levitt. Two photos introduced each artist's work. Participants started the information search on a start page with six links (number 1 in Figure 3) each connected to a short description of the corresponding artist (number 2; in the figure, the description of the first artist, Astrid Kruse-Jensen, is displayed.). Below the description, two further links were displayed (number 3), each of which led to one of the artist's photos (number 4). We made use of the Internet environment *http://picasa.com* to present the photos, tags and expert-generated keywords. Six of the twelve photos were presented with four tags appearing to the right (number 5) and four expert-generated keywords appearing below (number 6). The other six photos were presented with the tags appearing below and the keywords to the right.

On the whole, the application included twelve resources (photos), 48 tags and 48 expert-generated keywords. Note that we had drawn both the tags and the expert-generated keywords from the same word list in the following way: First, we gathered words in a pre-study where participants described the twelve photos with their own words. From this pool of words we then created two lists L1 and L2 so that they were equal with respect to printed frequency (M_{L1}=91.7, SD_{L1}=89.3; M_{L2}=91.4, SD_{L2}=60.2; $F_{1,94}$=0.0002, n.s.), concreteness (M_{L1}=3.55, SD_{L1}=2.18; M_{L2}=3.62, SD_{L2}=2.07; $F_{1,94}$=0.02, n.s.) and connectivity (M_{L1}=1.57, SD_{L1}=0.79; M_{L2}=1.43, SD_{L2}=0.74; $F_{1,94}$=0.88, n.s.). The latter three variables were determined by the word association norms of Nelson, McEvoy & Schreiber (1998a). Finally, for one half of the participants we drew the tags from L1 and the expert-generated keywords from L2; for the other half of the participants L1 served as keyword list and L2 as tag list.

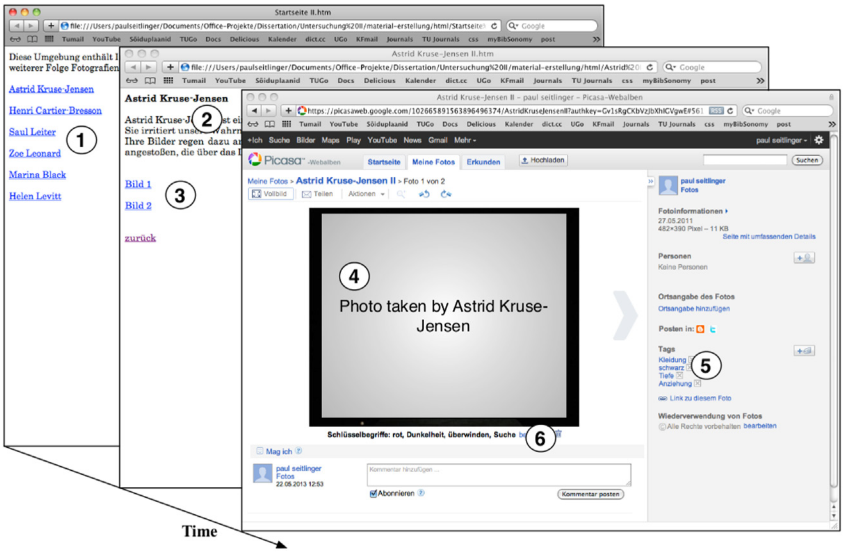

Figure 3. Application for information search in experiment 1
Note. Numbers described in the text

3.4.3 Search tasks and procedure

At the beginning, an instruction in written form asked the participant to search for decorative photos by means of the application run on a notebook (see Figure 3). Per each of the six artists, the participant read through the description, looked at both photos, the associated tags and keywords, and finally, selected one of the two photos of the same artist. The instruction led the participant to believe that the keywords had been created by experts in art history, and the tags had been created by participants working under similar conditions.

After that, the participant started conducting the search task (phase S). On average, going through all six artists and deciding on each of the six photo pairs took about ten minutes (M=9.62, SD=2.02).

Subsequently, the participant performed the distractor task, for which she/he was given a sheet of paper with 30 arithmetic problems (e.g. 729/70 = ?). The participant had to solve as many problems as possible within 60 seconds. As mentioned, the distractor task was performed to ensure participants were only drawing on secondary memory representations of the tags they had seen in the previous phase. Afterwards, the first test phase T_1 was conducted: all twelve photos of phase S were presented randomly on the notebook and the participant had to describe each photo on a second sheet of paper. The participant was asked to choose any words she/he deemed appropriate, either words that easily came to mind, or tags or keywords that had been experienced during phase S. Then the participant turned to the next photo by pressing a key. After the last photo presentation in T_1, the participant continued with the distractor task. The second test phase T_2 was followed by the third and last distractor task, which was then followed by T_3.

3.4.4 Design

The independent variables form a simple design constituted by the within-subjects factor word class (tag vs. keyword). The main dependent variables are the estimates of the parameters D, R, J_1, J_{2r}, J_{2n}, G.

With respect to the multinomial model analysis we have set α and β at a level, which allows for a compromise between the large sample size (N=2304; 24 participants × 96 tags) and the power of the statistical test (see Buchner, Erdfelder & Vaterrodt-Plünnecke, 1995). α and β are set equal since we deem both types of errors to be equally important. According to the effect size in case of multinomial modeling, a reasonable criterion is the detection of small effects (ω=.10 for χ^2-tests; see Buchner et al., 1995). We test the fit of the proposed model M_{2SDPg} consisting of two sub-models (one for tags and one for expert-generated keywords). In each sub-model, there are seven independent response categories (out of eight) and six free parameters. Hence, across the two experimental conditions there are 14 independent categories and 12 free parameters, resulting in two degrees of freedom. For the corresponding parameter values (ω=.10, N=2304 and df=2), the compromise power analysis (Erdfelder, Faul & Buchner, 1996) indicates the critical levels of $\alpha = \beta = 0.02$ (χ^2_{crit}=8.05)[1]. We apply the

---

[1] For the alternative models that contain only ten free parameters the critical values are set at χ^2_{krit}=10.79 corresponding to $\alpha = \beta = 0.03$.

software *MultiTree* (Moshagen, 2010) to test the multinomial model and to estimate the parameter values. As a second dependent variable, we analyze the number of repetitions of words from phase S across the three test phases T_i by means of an ANOVA with test phase (T_1 vs. T_2 vs. T_3) and word class (tag vs. keyword) as within-subjects factors. Given statistical error probabilities $\alpha = \beta = .05$ and the sample size N=24, we are able to detect an effect size of f=0.34.

3.4.5 Results and discussion

The raw frequencies that underlie the model-based analysis of experiment 1 can be found in the appendix (Table 11). To address our first research question, as to whether the proposed model accommodates the observed data, we test the 2SDPg's global fit to the full set of frequency data points. From the first row of Table 3, the 2SDPg test, it becomes clear that the goodness-of-fit measure, G^2, by no means exceeds the critical value. Hence, the 2SDPg proves to be compatible with participants' tag productions. To the contrary, the first two alternative models, 1SDPg and 2SOPg, have to be rejected since the corresponding goodness-of-fit measures exceed the critical value by a wide margin and are substantially worse than the fit of 2SDPg (see last column in Table 3). Additionally, the test of 2SDP, the original FTT-model, reveals that the inclusion of the guessing parameter G proves to be necessary to account for the data. While the simpler 2SDP-model's fit is barely below the critical value, its data fit is significantly worse than the fit of 2SDPg.

Insert Table 3 **about here**

Table 4 shows the parameter estimates and the corresponding 95% confidence intervals based on the 2SDPg. We first describe and discuss the probability estimates of tags shown in the first row. These estimates reveal that both a verbatim imitation based on direct retrieval (parameter *D*) and a semantic imitation based on reconstruction (parameter *R*) substantially contribute to participants' tag productions. With respect to the second step of semantic imitation, the familiarity judgment, we focus on the parameter J_1 since J_{2r} and J_{2n} are only included to account for participants' response tendencies in later test phases and, hence, are less important for our research questions. As Table 4 shows, J_1 is strongly pronounced and indicates a high probability of selecting a reconstructed word for output given the word has been studied in phase S.

Insert Table 4 **about here**

Referring to the question as to whether tags are encoded as strongly as expert-generated keywords (KWs), the two rows in Table 4 indicate quite similar parameter estimates for tags and KWs, especially if we consider the large overlap of the confidence intervals. The statistical test of the null hypothesis (e.g. $D_{(\text{Tags})}=D_{(\text{KWs})}$) is conducted by subtracting G^2 of the full model 2SDPg, i.e. 0.03, from G^2 of the sub model that is restricted by the corresponding null hypothesis. The difference between the full and sub model can be interpreted as χ^2 and its number of *df*s is the number of parameters by which the two models differ. If the difference exceeds the critical chi-square value, which is 3.84 in case of one *df*, the null hypothesis has to be rejected (for details about testing hypothesis by means of multinomial models see e.g. Batchelder & Riefer, 1990). With respect to the direct retrieval *D* and the reconstruction *R*, the restrictions $D_{(\text{Tags})}=D_{(\text{KWs})}$ and $R_{(\text{Tags})}=R_{(\text{KWs})}$ result in the model fits $G^2(3)=0.33$ and $G^2(3)=0.25$, respectively, which differ from the full model by $\chi^2(1) = 0.30$ and $\chi^2(1) = 0.22$, respectively. Since both differences are not significant, the two word categories appear to leave comparably strong verbatim and gist traces during the information search. Similarly, imposing the restrictions $J_{1(\text{Tags})}= J_{1(\text{KWs})}$ and $G_{(\text{Tags})}=G_{(\text{KWs})}$ results in nested models that account for the data. The corresponding model fits are $G^2(3)=0.08$ and

$G^2(3)$=0.12, which differ from the full model by $\chi^2(1) = 0.05$ and $\chi^2(1) = 0.09$, respectively. Clearly, these two differences are not significant.

The 3 (test phase: T_1 vs. T_2 vs. T_3) × 2 (word class: tags vs. keywords) repeated measures ANOVA on the number of repetitions reveals no significant effects neither for the two factors word class and test phase nor for an interaction (all Fs < 1).

Summing up, these results provide evidence for the first hypothesis 1.1 that the proposed fuzzy-trace model of ST accommodates the probability of tag imitations in the STTT procedure. First, the statistical test has yielded a very small divergence between the data observed and the data predicted by the 2SDPg model. Second, the fit of this model has turned out to be substantially better than the fit of three alternative models. Furthermore, our parameter estimates show that both the semantic and verbatim process has a substantial influence on the imitation of previously seen tags: Since 2SDPg has strongly outperformed 1SDPg we conclude that the semantic process involves a reconstruction operation and a subsequent familiarity-based choice of tags.

With respect to our second hypothesis 1.2, the encoding of user-generated tags corresponds in strength to the encoding of expert-generated keywords in phase S. Thus, both expert-generated keywords and user-generated tags seem to be relevant cues during the information search.

To quantify the extent to which verbatim imitation and semantic imitation are involved in the imitation of tags (research question RQ2), it is possible to decompose the relative frequency of tags that have been reused in T_1 by using the parameter estimates of D, R, J_1 and G into verbatim and semantic imitation. The relative frequency of reusing a tag in T_1 amounts to p=.29[2]. The first column of Table 5 shows that this relative frequency can be approximately

---

[2] The frequency of p=0.29 is easily calculated by dividing 338, which is the sum of the frequencies of response categories including a tag reuse in T_1, i.e., rrr, rrn, rnr, rnn, by 1152, which is the sum of the

explained by adding up the estimated probabilities of verbatim imitation, semantic imitation and no imitation leading to t_r. The estimates result from inserting the estimates of the parameters D, R, J_1 and G of the first row of Table 4 into the model equations of Table 2 (second column).

Insert Table 5 **about here**

The row sums of Table 5 reveal that both verbatim and semantic imitation substantially contribute to the participants' tag productions in test phase T_1. While the proportion of tags produced via verbatim imitation is estimated at about 20%, the proportion of tags produced via semantic imitation is about 13%.

Finally, since we have taken the measurement model from the domain of memory psychology, we cannot be sure that the parameter interpretations are entirely valid within the domain of tagging photos. Therefore, we conclude this section by describing a pattern of results that speaks in favor of the validity of our interpretations. According to FTT, recall from memory based on verbatim traces proceeds faster than gist-based reconstruction (e.g. Brainerd et al., 2002). Consequently, we would assume that tags applied to a picture in an early output position would more likely be results of recall from verbatim traces. Similarly, a central assumption of the STTT procedure and its measurement model is that words, which are recalled in each test phase (i.e., belong to the response category rrr), are stored in form of verbatim traces. We can therefore predict that within the category rrr (in the following denoted as "verbatim response category"), early positions of the used tags should occur more frequently than later output positions. To the contrary, within the remaining set of reused tags

---

frequencies of all response categories for tags. The raw frequencies are presented in the appendix in the first data column of Table 11.

(in the following denoted as “semantic response category”) this relationship should be reversed. Thus, we predict an association between the category of reused tags and the output position.

To compare this prediction with the data, we looked at the sequence of tags produced by each participant in each of the three test phases. Since only a small portion of participants assigned more than 4 tags to a photo (4.2 %), we restricted the analysis to the first four output positions per photo. Then, for each participant, we counted the number of times the two response categories (verbatim vs. semantic) were associated with one of the four output positions.

Insert Table 6 about here

Table 6 provides the results in form of a 2 (response category: “verbatim response category” vs. “semantic response category”) × 4 (Output position: positions 1 to 4) contingency table. The pattern of these relative frequencies seems to be well in accordance with our prediction. Among the “verbatim response category” tags, the relative frequencies decrease monotonically as a function of the output position. To the contrary, among the “semantic response category” tags, the relative frequencies increase as the output position gets larger. We have also performed a 2 (“verbatim response category” vs. “semantic response category”) × 4 (positions 1 to 4) χ^2-test[3], which shows the difference between observed and expected values to be large and significantly higher than expected under the null hypothesis of no association ($\chi^2(3) = 23.5$, $p < .01$). Hence, we conclude that some association exists between output position and memory processes in tag production where the contributions of “verbatim response category” tags get increasingly smaller as participants continuously reuse

[3] The absolute frequencies underlying the χ^2-test can be easily calculated by multiplying the relative frequencies by the corresponding row sums, which are 808 (“verbatim response category”) and 165 (“semantic response category”).

tags to describe a particular photo. To the contrary, tags of the “semantic response category” tend to occur on later output positions. As this result is in line with our predictions derived from FTT, we assume the estimated parameters can be interpreted in terms of verbatim and gist-based processes, and support fast and slow recall from memory, respectively. Further results speaking in favor of our parameter interpretations are provided in experiment 2 where we introduced and manipulated independent variables that are assumed to differentially affect the corresponding parameters.

3.5 Experiment 2

The aims of the second experiment are threefold. First, we replicate the results of the first experiment in an independent sample of participants. Second, we realize more natural conditions during the STTT-procedure that should allow for generalizing our conclusions. Third, we examine the question, as to whether particular design features of an ST-environment (such as the size of displayed tags) influence memory processes measured by the parameters *D*, *R* and *J*.

The first hypothesis is that the proposed model also accounts for the data observed in an independent sample under less controlled conditions (hypothesis 2.1). For this purpose, we realized the STTT-procedure in an online environment and all subjects participated from home or from their workplace instead of under laboratory conditions.

With respect to the research question RQ3 as to whether particular design features influence the parameters *D*, *R* and *J*, we draw on dual-process memory research (e.g. Yonelinas, 2002) and vary three variables. With respect to the process represented by *D*, Brainerd & Reyna (2010) regard the direct retrieval *D* as a recollective operation. According to the dual-process research, recollection requires elaborative encoding of the learning material that is usually defined as reflecting and recognizing relations to preexisting knowledge. Strongly related to this interpretation of elaborative activity is the connectivity of a word, a cognitive construct examined by Nelson and colleagues (e.g. Nelson et al., 1998b; Nelson, Goodmon & Akirmak, 2007). Operationalized by means of extensive association norms (e.g. Nelson et al., 1998a), the connectivity is an index of the extent to which the associations of a word's memory representation are in turn associated to each other. *In the context of ST we therefore assume that tags scoring high in connectivity leave stronger memory representations and thus have a higher probability D being directly retrieved during a tag assignment than tags scoring low in connectivity (hypothesis 2.2)*.

With respect to the parameter *R*, the probability of the reconstruction operation, semantic organization of a word list increases its estimates (e.g. Brainerd & Reyna, 2010). In the context of ST, Schrammel, Leitner and Tscheligi (2009) have found that semantically organized tag clouds accelerate users' search processes. In this work, we examine whether they also have a positive impact on the gist-based reconstruction. *We assume that semantically organized tag clouds result in larger estimates of the parameter R than randomly organized tag clouds (hypothesis 2.3).*

Finally, encoding manipulations, such as duration and attention, influence the familiarity-based retrieval of words, i.e. *J*. In the context of ST, large tags draw users' attention more strongly than small tags, resulting in longer reading times during information search (Schrammel et al., 2009). With respect to the underlying memory processes, we assume that after reconstruction, large tags evoke a stronger feeling of familiarity and are chosen with a higher probability than small tags. *Hence, we hypothesize that large tags result in larger estimates of the parameter J than small tags (hypothesis 2.4)*.

3.5.1 Participants

48 subjects (36 female) with an average age of 25.9 years (*SD*=12.6, ranging from 16 to 53 years) participated in the experiment for a 10€ cinema coupon. They were recruited from two different university institutes (Department of Psychology at University of Graz and

Knowledge Technologies Institute at Graz University of Technology) and equally assigned to the two experimental groups.

3.5.2 Materials

Environment. The STTT-procedure was realized using an online survey environment created by means of *LimeSurvey* (*http://limesurvey.org*). The participants received a link to the environment, in which they performed the STTT-procedure. The learning material was the same as in experiment 1. After a short instruction, a first, randomly drawn artist description appeared. A subsequent site presented the artist's two photos (see number 1 in Figure 4).

A tag cloud (number 2) including 16 tags, eight in large and eight in small type size, described each photo. Below, the participant declared her decision by selecting either photo 1 or photo 2 (number 3 in Figure 4). The remaining five artists were presented in the same way. On the whole, the environment included twelve photos (two per artist), twelve tag clouds (one per photo) and 192 tags (16 per tag cloud). For each photo, two different kinds of tag clouds were used, one with randomly structured tag clouds (random layout condition) and one with semantically structured tag clouds (semantic layout condition). In the latter layout condition, semantically similar tags were positioned next to each other.

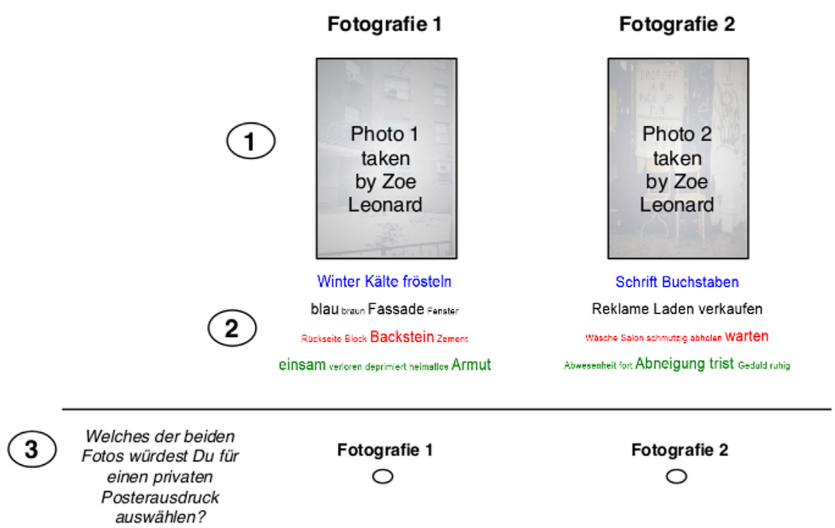

Figure 4. Presentation of photos and tag clouds in experiment 2

Note. The question (at number 3) to be answered by selecting photo 1 (Fotografie 1) or photo 2 (Fotografie 2) means "Which of the two photos would you choose for a private poster print?"

Tag clouds. Each of the twelve tag clouds was made up of eight tags in large (20 pt) and eight tags in small (12 pt) type size where each size set was equally split into tags with high and low connectivity values (determined from the word association norms of Nelson et al., 1998a). This 2×2-variation resulted in the four conditions "large size/high connectivity", "large size/low connectivity", "small size/high connectivity" and "small size/low connectivity". To realize each of the four conditions we had created four lists by drawing words from the pool generated in the pre-study of experiment 1. The two lists L1 and L2 were high in connectivity (L1: M=2.03, SD=0.53; L2: M=2.00, SD=0.41) and the two lists L3 and L4 were low in connectivity (L3: M=1.03, SD=0.40; L4: M=0.98, SD=0.32). The four lists were equal with respect to concreteness ($F_{3,187}$=0.31, n.s.) and printed frequency ($F_{3,187}$=0.02, n.s.). Each of the four lists was used for one of the four *type size* × *connectivity* conditions. Finally, to create semantic tag clouds, we adopted the procedure suggested by Schrammel et al. (2009). First, we determined the similarities between each pair of the 16 tags of a tag cloud by LSA (*http://lsa.colorado.edu/*) and represented them in a matrix. Second, by performing multidimensional scaling (ALSCAL) we transformed the matrix into a two-dimensional space, in which semantically related tags (e.g. cold, winter, freeze) are represented in clusters that are remote from semantically dissimilar clusters (e.g. block, brick, cement). Finally, we arranged the words within a cluster in one common row of a tag cloud.

3.5.3 Procedure

On average, an individual session took 39 minutes (M=39.1, SD=15.7). Each participant received a link to the environment for either the random or the semantic layout condition. Hence, layout was a between-subjects factor. After a short introduction, the artist descriptions were presented on subsequent pages, each description followed by the two associated photos and tag clouds (see Figure 4). After the sixth artist, the participant had to perform the first distractor task. Then, all twelve photos of phase S were presented on one single page, and the participant was asked to enter her/his tags in an empty field below each photo. The participant was asked to choose any words she/he deemed appropriate, either words that easily came to mind, or tags that had been experienced during phase S. After the participant had described each of the twelve photos, she/he performed the second distractor task. Then, the second test phase T_2, the third distractor task and finally, the third test phase T_3 were administered. On a last page the participant was debriefed about the experiment's goals.

3.5.4 Design

The independent variables form a mixed 2 × 2 × 2-Design constituted by the factors a) type size (large vs. small, within-subjects), b) connectivity (high vs. low, within-subjects) and c) layout (semantic vs. random, between subjects). The main dependent variables are again the parameters of the multinomial processing tree DP2Sg, i.e., D, R, J_1, J_{2r}, J_{2n}, G. Since in the context of social tagging, studies on our specific questions do not exist and we cannot exclude very small effects of the independent variables, we decided to switch to sensitivity power analysis with respect to the multinomial model analysis. Given statistical error probabilities $\alpha = \beta = .05$ and the sample size N=4608 (24 participants per group × 192 tags) and df=8 (56 free categories − 48 free parameters), we are able to detect a very small effect of size f=0.07 (χ^2_{crit}=15.51). As a second dependent variable, we analyze the number of repetitions of tags from phase S across the three test phases T_i by means of an ANOVA with test phase, type size and connectivity as within-subjects factors and layout as between-subjects factor. Given statistical error probabilities $\alpha = \beta = .05$ and the sample size N=48, we are able to detect an effect of size f=0.17.

3.5.5 Results and discussion

The raw frequencies that underlie the model-based analysis of experiment 2 are presented in the appendix (see Table 12 for the semantic layout and Table 13 for the random layout condition). The analysis reveals that the proposed model, the 2SDPg, fits the entire set of frequency data well. In fact, the divergence between predicted and empirical data proves to be small, $G^2(8)$=2.78, and far below the critical value, χ^2_{crit}=15.51.

Table 7 shows the estimates of the parameters *D*, *R*, J_1 and *G* for each condition as well as the means of each type of these four parameters in boldface (the entire set of parameter estimates and the corresponding 95% confidence intervals is given in the appendix in Table 14). Overall, the table provides evidence for hypothesis 2.1 since the pattern of estimates resembles the pattern from experiment 1. The average estimates of *D*, *R* and J_1 and *G* are .08, .14 and .47, respectively. Hence, both the direct retrieval and reconstruction operation substantially contribute to the tag imitations and the estimates of the familiarity judgment are strongly pronounced.

In accordance with hypothesis 2.2, high connectivity tags evoke larger estimates of the *D* parameter than low connectivity tags. This effect appears to be stronger under the random layout condition. In the random layout condition group, the fit of the full model without any restrictions is $G^2(2)$=0.31, whereas a sub model constrained by equal *D*-parameters for high and low connectivity tags yields a goodness-of-fit of $G^2(3)$=4.50. The difference to the full model is $\chi^2(1)$=4.19 and exceeds the critical value of 3.84. Under the semantic layout condition, on the other hand, the goodness-of-fit measures for the full and the sub model with corresponding *D* parameter restrictions are $G^2(2)$=1.06 and $G^2(3)$=2.89, which do not differ significantly.

As expected, the *R* parameters are influenced by the tag cloud layout. In fact, setting the *R* parameter constant across the two groups yields a model fit of $G^2(12)$=14.92 that differs from the full model (see above) by $\chi^2(4)$=12.14. This difference is greater than the critical value of

9.49. In contrast to what we had expected in hypothesis 2.3, however, the estimates are larger under the random than under the semantic layout condition.

In line with hypothesis 2.4, tags presented in large type size seem to evoke larger estimates of the parameter J_1 than tags presented in small type size. Again, this effect appears to be influenced by the tag cloud layout. In the random group, the full model with non-restricted J_1 parameters yields a model fit of $G^2(2)=0.08$; a sub model constrained by equal J_1 parameters for large and small type size yields $G^2(3)=4.83$. The difference between both models is $\chi^2(1)=4.74$ and greater than the critical chi-square value of 4.19. In the semantic group, the J_1-parameter is not influenced by the type size. Imposing the restriction J_1(large)=J_1(small) does not reduce the model fit significantly, $G^2(3)=1.25$.

The 3 (test phase: T_1 vs. T_2 vs. T_3) × 2 (type size: large vs. small) × 2 (connectivity: high vs. low) × 2 (layout: semantic vs. random) Mixed ANOVA on the number of repetitions of tags reveals highly significant effects for type size ($F_{1,46}=20.63$, $p<.001$) and connectivity ($F_{1,46}=29.20$, $p<.0001$) but no significant results for the other two factors (test phase and layout) as well as for any possible interaction (all Fs<1).

Insert Table 7 **about here**

As in experiment 1, the values of the parameters D, R, J_1 and G can be inserted into the model equations of Table 2 to decompose the relative frequency of reused tags in T_1 into verbatim and semantic imitations. Here, we have decomposed the frequency of reused tags in T_1 under the semantic and random layout condition, which are 0.14 (semantic layout) and 0.16 (random layout)[4] on the basis of the parameter means (presented in boldface in Table 7).

---

[4] See experiment 1 for calculating these relative frequencies. The raw frequencies are presented in Table 12 and Table 13 in the appendix for the semantic and random layout condition, respectively.

The first and fourth column sum of Table 8 show that both relative frequencies can be explained well by summing the estimates of verbatim imitation, semantic imitation and no imitation (resulting in t_r). Considering the means of the corresponding row sums, the verbatim contribution to tag productions amounts to 8%, the semantic contribution to 13% on average. As the row sums of Table 8 show, the difference between the two sorts of imitation depend on the layout condition because – as described above – the parameter *R* varies with this variable. While the contribution of verbatim imitation is similar under the semantic and random layout condition (about 8%), semantic imitation is more strongly pronounced under the random (about 16%) than under the semantic layout condition (about 9%).

Insert Table 8 **about here**

Because of this difference of semantic imitation in the two groups (semantic vs. random layout), we can test one further assumption that speaks in favor of the validity of our interpretation of semantic imitation[5]. If the reconstructive process measured by the parameter *R* really leads to semantic imitation, then the new tags produced as a result of semantic imitation should be semantically more similar to the original tags than those produced under a condition of less semantic imitation.

To test this prediction empirically, we have calculated LSA scores between the new tags assigned by a participant and the 32 tags of the corresponding tag clouds for each participant and each artist by using One-To-Many Comparison (*http://lsa.colorado.edu/*). Table 9 provides the means and standard deviations of these LSA scores separated by tag cloud layout

---

[5] We thank one of the reviewers of an earlier version of the manuscript for bringing this to our attention.

and artist. The table reveals small but consistent differences between the two groups in favor of the random condition that seem to be independent of the artist.

To test the differences for statistical significance, we have performed a 2 (tag cloud layout: random vs. semantic) × 6 (1st to 6th artist) Mixed ANOVA on the LSA scores. The factor "tag cloud layout" is significant at the .05 alpha level ($F_{1,36}=4.53$, $p<.05$) indicating that the tags produced under the random layout condition are semantically more similar to the original tags in the tag clouds than the tags produced under the semantic layout condition. The ANOVA does not reveal significant effects for the factor artist and for an interaction (all *F*s<1).

This result corresponds with our expectation that a group exhibiting higher estimates of semantic imitation produces new tags, which are more semantically similar to presented tags than new tags produced by a group exhibiting lower estimates of semantic imitation. Thus, we regard this result as supporting our interpretation of semantic imitation as a gist-based reconstruction.

Insert Table 9 **about here**

Overall, the results of experiment 2 are well in accordance with hypothesis 2.1 since the proposed model DP2Sg provides a good account of the data observed both under the random and semantic layout condition. Furthermore, the results provide evidence for the hypotheses that connectivity influences the parameter D (hypothesis 2.2), that the tag cloud layout influences the parameter R (hypothesis 2.3) and that type size influences the parameter J_1 (hypothesis 2.4). With respect to hypothesis 2.3, however, we have obtained the counterintuitive result that randomly – and not semantically – arranged tags increase the estimates of parameter R. Although further experiments are necessary to gather additional evidence, we assume that a random arrangement impedes fast and automatic processing and stimulates a meaningful reflection of the tag cloud. If participants notice that some tags in the cloud belong to a common semantic category, they might actively group them, leading to a generation effect (e.g., Slamecka & Katsaiti, 1987). The consequence might be stronger gist traces of encountered tags and thus, larger estimates of the parameter *R*.

From these results, we conclude that verbatim and semantic imitation contribute independently to the participants' tag assignments. This is because the connectivity of tags, the tag cloud layout and the type size of tags appear to have specific effects on the parameters *D*, *R* and J_1, respectively. According to Buchner et al. (1995), specific effects of experimental variables on model parameters indicate a functional independence of the processes represented by the parameters.

4 General discussion

We have conducted the two experiments to pit models of verbatim and semantic imitation against each other and to dissociate the contributions of verbatim and semantic processes to the probability of a tag being reused, t_r. In particular, we tested the assumption that two qualitatively distinct memory representations are in play during the imitation, a verbatim and a gist trace of previously seen tags. While the verbatim trace should support a direct retrieval and – in any case – t_r, the gist trace allows for a reconstruction operation resulting in t_r only in case of a positive familiarity judgment. Both experiments confirm these assumptions and reveal that the two mechanisms, the direct retrieval of a tag's surface form and the reconstruction of its meaning, have substantial portion in the generation of tags.

Moreover, a comparison of the first and second experiment shows that the contribution of the gist trace to the memory-based imitation of tags is more robust against different conditions than the contribution of the verbatim trace. While the probability of a direct retrieval goes down from 0.20 (experiment 1) to 0.08 (experiment 2), the probability of a gist-based reconstruction remains constant at about 0.13. The reason for this differential decline is that in experiment 2 twice as many tags have been presented during the study phase as in experiment 1. This pattern of results is in accordance with FTT that assumes verbatim traces to be more prone to proactive interference, for instance caused by subsequent verbal stimuli, than gist traces (e.g. Brainerd & Reyna, 2010). This provides further evidence for the validity of the model.

Summarizing from these results and additional analyses, we have received very strong evidence across both experiments that our interpretations of model parameters correspond to the original interpretation of FTT, namely the specific effects of the independent variables in experiment 2 and the differential robustness of verbatim and gist traces. In addition to that, an analysis of experiment 1 has clearly demonstrated that tags of verbatim response categories tend to appear in early positions, while tags of semantic response categories appear in late output positions. Hence, also in the domain of tagging, the parameters *D* and *R* seem to represent processes that differ in speed of recall. Finally, we have observed that differences in the estimates of the parameter *R* are associated with corresponding differences in the LSA-scores of newly introduced tags. This last result speaks in favor of the assumption that reconstructive processes underlie semantic tag imitation.

Our study also demonstrates that modeling memory processes involved in semantic imitation allows for a more in-depth analysis of the underlying processes. The counterintuitive result of the tag cloud layout would not have been revealed without this approach. Now, a deeper analysis of exactly which layout features contribute to a higher versus lower amount of semantic elaboration can be conducted.

Additionally, we deem the present work important for the development of recommendation mechanisms supporting tag assignments (e.g. Jäschke et al., 2008) in social information systems. For instance, in a recent work, Seitlinger, Kowald, Trattner and Ley (2013) have emulated the gist-based production of tags as a simple connectionist model and demonstrated that a tag recommender based on this model provides more accurate recommendations than a well-established Latent Dirichlet Allocation approach

A limitation of our study is that in both experiments our focus was on (long-term) secondary memory (e.g. James, 1890; Craik, 1971; Unsworth & Engle, 2007), in particular, on durable memory traces of previously seen tags supporting episodic memory during tag assignments. This is why participants were required to perform the distractor task before the test phases. The reason for emphasizing secondary memory processes was that they have been found to play a major role in ST (e.g. Fu, 2008; Fu & Dong, 2012). . In our future work, we will more specifically focus on the impact of primary memory as a trigger for searching secondary memory by cues from external environment (e.g. Unsworth & Engle, 2007). We will also examine how this process relates to particular design choices in the user interface, such as tags from recommendation services displayed while tags are assigned.

A further limitation is that our research design and measurement model do not take into account different motivations for tagging Web resources. The specific setting realized in experiment 1 and 2 and the parameter estimates obtained may generalize to situations, in which users tag for the purpose of future retrieval and sharing. If users are motivated to contribute to personal and collaborative information organization, they may be inclined to verbalize resource-specific, semantic aspects and also to adopt existing tags. However, the parameter estimates will probably change, if tagging is used for social signaling and self-presentation. Under such motivational conditions, users may intend to invent original tags to attract attention (e.g. Marlow et al., 2006). While gist-based and thus, to some extent implicit memory processes (see Brainerd & Reyna, 2010) may still contribute to the invention of new tags, verbatim imitation will be quite unlikely. Therefore, future studies that experimentally vary users' intentions and aim at measuring this variation by means of an additional, "originality" parameter will help investigating the impact of users' intentions on verbatim and semantic imitation.

Moreover, by using photos as Web resources, we have focused on a specific format, which helped reduce the influence of additional, verbal cues. We must leave the question open for future studies as to whether attention and encoding of tags vary across different formats (e.g., photos vs. texts) and whether format-specific factors determine the relative mix of verbatim and semantic imitation. Several explorative model-based analyses we conducted with the present data seem to indicate larger estimates of the parameter *D* for concrete (e.g. "street") than for abstract tags (e.g. “patience”), but no differences in the estimates of the parameter *R*. In the case of photos, concrete tags may be encoded both verbally and visually by drawing the user's attention to particular elements of the photo, and we assume that such dual encoding (Paivio, 1986) leads to stronger contributions of verbatim imitation (i.e., higher estimates of the parameter *D*) than in case of abstract tags. .

In case of different formats (texts, audio, etc.), neither concrete nor abstract tags can refer to pictured objects and thus, dual coding is less likely to occur. In the case of texts, additional verbal cues (e.g., in title or abstract) may evoke stronger semantic processing resulting in a stronger contribution of gist-based reconstruction. Future studies may therefore investigate the question as to whether the impact of factors, such as concreteness or semantic narrowness of tags (e.g., “oak” vs. “plant”), on the relative contributions of verbatim and semantic imitation depends on the resource's format.

Finally, our results reveal positive pedagogical implications of ST for cooperative and collaborative learning settings. The estimates of the parameter *R* prove to be substantial and robust. The parameter measures the gist-based reconstruction of meaning, which goes beyond the word level and concerns less the tags but rather the thoughts and concepts underlying the tags. In line with Fu et al. (2010) we therefore conclude that ST is a phenomenon to be examined both on a word and semantic level. Particularly, social tags are not only cues supporting information search or leading to simple swarm-like imitation (Ley & Seitlinger, 2011), but seem to externalize individual thoughts that in turn evoke associated ideas and reflections of other students. Considered in this light, it becomes clearer why the simple tagging functionality of collaborative learning environments proves to support learning processes among students (e.g. Kuhn et al., 2012). As a tool that interconnects both external (e.g. learning objects) and internal representations (e.g. gist traces), ST has the potential to foster the exchange of learning experiences in constructivist and technology-enhanced learning settings.

5 References

Batchelder, W. H., & Riefer, D. M (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548-564. DOI: 10.1037/0033-295X.97.4.548

Batchelder, W. H., & Riefer, D. M (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57-86. DOI: 10.3758/BF03210812

Bates, M. J. (1998). Indexing and access for digital libraries and the internet: human, database, and domain factors. *Journal of the American Society for Information Science, 49*, 1185-1205. DOI: 10.1002/(SICI)1097-4571(1998110)49:13<1185::AID-ASI6>3.0.CO;2-V

Bertrand, A., Cellier, J.-M., & Giroux, L. (1996). Expertise and strategies for the identification of the main ideas in document indexing. *Applied Cognitive Psychology, 10*, 419-433. DOI: 10.1002/(SICI)1099-0720(199610)10:5<419::AID-ACP401>3.0.CO;2-W

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780195154054.001.0001

Brainerd, C. J., & Reyna, V. F. (2010). Recollective and non-recollective recall. *Journal of Memory and Language, 63*, 425-445. DOI: 10.1016/j.jml.2010.05.002

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*, 120-152. DOI: 10.1006/jmla.2001.2796

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124*, 137-160. DOI: 10.1037/0096-3445.124.2.137

Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic Dynamics and Collaborative Tagging. *Proceedings of the National Academy of Sciences, 104*, 1461-1464. DOI: 10.1073/pnas.0610487104

Craik, F. I. M. (1971). Primary memory. *British Medical Bulletin, 27,* 232–236.

David, C., Giroux, L., Bertrand-Gastaldy, S., & Lanteigne, D. (1995). Indexing as problem solving: a cognitive approach to consistency. In T. Kinney (Ed.), *Proceedings of the American Society for Information Science* (pp. 49-55). Medford, NJ: Information Today.

Dellschaft, K., & Staab, S. (2008). An epistemic dynamic model for tagging systems. In P. Brusilovsky, & H.C. Davis (Eds.), *Proceedings of the 19th ACM conference on Hypertext and Hypermedia* – Hypertext 2008 (pp. 71-80). New York: ACM-Press. DOI: 10.1145/1379092.1379109

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagueren, M., & Nadarevic, L. (2009). Multinomial processing tree models. A review of the literature. *Journal of Psychology, 217*, 108-124. DOI: 10.1027/0044-3409.217.3.108

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28*, 1-11. DOI: 10.3758/BF03203630

Farrow, J. (1995). All in the mind: concept analysis in indexing. *The Indexer, 19*, 243-247.

Fu, W.-T., & Dong, W. (2012). Collaborative indexing and knowledge exploration: a social learning model. *Intelligent Systems, 27*, 39-46. DOI: 10.1109/MIS.2010.131

Fu, W.-T., Kannampallil, T. G., & Kang, R. (2009). A semantic imitation model of social tag choices. In *Proceedings of the 2009 International Conference on Computational Science and Engineering – CSE'09* (pp. 66-73). New York: ACM-Press. DOI: 10.1109/CSE.2009.382

Fu, W.-T., Kannampallil, T. G., Kang, R. & He, J. (2010). Semantic imitation in social tag-ging. *ACM Transactions on Computer-Human Interaction, 17*, 12:1-12:37. DOI: 10.1145/1806923.1806926

Gerjets, P. & Kammerer, Y. (2010). Topical relevance and information quality in cognitive models of Web search behavior: Introducing epistemic scent into information foraging theory. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2278-2283). Austin, TX: Cognitive Science Society.

Glushko, R. J., Maglio, P. P. Matlock, T., & Barsalou, L. W. (2008). Categorization in the wild. *Trends in Cognitive Science, 12*, 129-135. DOI:10.1016/j.tics.2008.01.007

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*, 198-208. DOI: 10.1177/0165551506062337

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition, 118*, 359-376. DOI: 10.1016/j.cognition.2010.10.009

Gomes, C. F. A., Brainerd, C. J., & Stein, L. M. (2013). The effects of emotional valence and arousal on recollective and nonrecollective recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 663-677. DOI: 10.1037/a0028578

Halpin, H., Robu, V., & Shepherd, H. (2007). The Complex Dynamics of Collaborative Tag-ging. In C. Williamson, M. E. Zurko, P. Patel-Schneider, & P. Shenoy (Eds.), *Proceedings* *of the 16th international conference on World Wide Web – WWW'07* (pp. 211-220). ACM-Press. DOI: 10.1145/1242572.1242602

James, W. (1890). *Principles of psychology.* New York: Holt.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2008). Tag recommendations in social bookmarking systems. *AI Communications, 21*, 231-247. DOI: 10.3233/AIC-2008-0438

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory*. New York: Academic Press.

Kuhn, A., McNally, B., Schmoll, S., Cahill, C., Lo, W., Quintana, C., & Delen, I. (2012). How students find, evaluate, and utilize peer-collected annotated multimedia data in science inquiry with Zydeco. In J. A. Konstan, E. H. Chi, & K. Höök (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '12* (pp. 3061-3070). New York: ACM-press. DOI: 10.1145/2207676.2208719

Ley, T., & Seitlinger, P. (2011). *Why people don't swarm: evidence for a dual-process memory model in collaborative tagging*. Poster presented at the 3rd International Conference on Web Science WebSci'11, Koblenz, Germany.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In U. K. Wiil, P. J. Nürnberg, & J. Rubart (Eds.), *Proceedings of the seventeenth conference on Hypertext and Hypermedia – HYPERTEXT'06* (pp.31-40). New York: ACM-press. DOI: 10.1145/1149941.1149949

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42-54. DOI: 10.3758/BRM.42.1.42

Nelson, D. L., Goodmon, L. B., & Akirmak, U. (2007). Implicitly activated memories are associated to general context cues. *Memory & Cognition, 35*, 1878-1891. DOI: 10.3758/BF03192922

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998a). The University of South Florida word association, rhyme, and word fragment norms. http://web.usf.edu/FreeAssociation/.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998b). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105*, 299-324. DOI: 10.1037/0033-295X.105.2.299

Nov, O., Naaman, M., & Ye, C. (2008). Analysis of participation in an online photo-sharing community: a multidimensional perspective. Journal of American Society for Information Science and Technology, 61, 555-566. DOI: 10.1002/asi.v61:3

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318-339. DOI: 10.1037/0033-295X.95.3.318

Sakamoto, Y., Ma, J., & Nickerson, J. V. (2009). 2377 people like this article: The influence of other's decisions on yours. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1959-1965). Amsterdam, Netherlands: Cognitive Science Society.

Schrammel, J., Leitner, M., & Tscheligi, M. (2009). Semantically structures tag clouds: an empirical evaluation of clustered presentation approaches. In D. R. Olsen, R. B. Arthur, K. Hinckley, M. R. Morris, S. Hudson, & S. Greenberg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI'09* (pp. 2037-2040). ACM-Press. DOI: 10.1145/1518701.1519010

Seitlinger, P., Kowald, D., Trattner, C., & Ley, T. (2013). Recommending tags with a model of human categorization. In Q. He, A. Iyengar, W. Nejdl, J. Pei, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM international conference on information and knowledge management – CIKM'13* (pp. 2381-2386). ACM-press. DOI: 10.1145/2505515.2505625

Seitlinger, P., & Ley, T. (2012). Implicit imitation in social tagging: familiarity and semantic reconstruction. In J. A. Konstan, E. H. Chi, & K. Höök (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI'12* (pp. 1631-1640). ACM-press.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., & Riedl, J. (2006). Tagging, communities, vocabulary, evaluation. In P. Hinds, & D. Martin (Eds.), *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work – CSCW'06* (pp. 181- 190). ACM-press. 10.1145/1180875.1180904

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26*, 589-607. DOI: 10.1016/0749-596X(87)90104-5

Stahl, C., & Klauer, K. C. (2007). A simplified conjoint recognition paradigm for the measurement of gist and verbatim memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 570-586. DOI: 10.1037/0278-7393.34.3.570

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104-132. DOI: 10.1037/0033-295X.114.1.104

Van Maanen, L., & Marewski, J. N. (2009). Recommender systems for literature selection: a competition between decision making and memory models. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2914-2919). Austin, TX: Cognitive Science Society.

Voss, J. (2007). Tagging, folksonomy & co – renaissance of manual indexing. 10th international Symposium for Information Science. In A. Oßwald, M. Stempfhuber & C.

Wolff (Eds.), *Proceedings of 10th international symposium of information science* (pp. 243-254). Konstanz: Universitätsverlag.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: an integrative review of the empirical literature. *Psychonomic Bulletin & Review, 1*, 89-106. DOI: 10.3758/BF03200763

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language, 46*, 441-517. DOI: 10.1006/jmla.2002.2864

Table 1. Latent processes underlying the categories of reusing/not reusing tags

		Observed category	
		Tag reused (t_r)	Tag not reused (t_n)
Latent process	Verbatim imitation	p(t_r|verbatim)	0
	Semantic imitation	$p(t_r \mid semantic)$	$p(t_n \mid semantic)$
	No imitation	$p(t_r \mid idiosyncratic)$	$p(t_n \mid idiosyncratic)$

Note. t_r=tag reused, t_n=tag not reused

Table 2. Relation between FTT-constructs and different types of tag-imitations

Latent			Manifest
Recall state	Probability of entering recall state	Type of imitation	Observed category
L	D	Verbatim imitation	t_r
P_r	$(1-D)RJ$	Semantic imitation	t_r
P_n	$(1-D)R(1-J)$	Semantic imitation	t_n
U_r	$(1-D)(1-R)G$	No imitation	t_r
U_n	$(1-D)(1-R)(1-G)$	No imitation	t_n

Note. t_r=tag reused, t_n=tag not reused, L=*Perfect recall* state, P_r=*Partial recall* state that leads to t_r, P_n=*Partial recall* state that leads to t_n, U_r=*No recall* state that leads to t_r, U_n=*No recall* state that leads to t_n, D=Direct retrieval, R=Reconstruction, J=Familiarity-based judgment, G= Guessing

Table 3. Results of model tests in experiment 1

Null hypotheses	G^2	*df*	χ^2_{krit}	
2SDPg test	0.03	2	8.05	
Alternative model tests				Difference to DP2Sg's fit
1SDPg	88.94	4	10.79	$\chi^2(2) = 88.91$***
2SOPg	197.70	4	10.79	$\chi^2(2) = 197.67$***
2SDP	9.72	4	10.79	$\chi^2(2) = 9.69$*

Note. **p*<.05, ****p*<.001

2SDPg=2-step dual process plus guessing, 1SDPg=1-step dual process plus guessing, 2SOPg=2-step one process plus guessing, 2SDP=2-step dual process

Table 4. Experiment 1 parameter estimates based on DP2Sg for tags and expert-generated keywords

	D	R	J_1	J_{2r}	J_{2n}	G
Tags	0.20 [0.15, 0.24]	0.16 [0.10, 0.22]	0.60 [0.44, 0.75]	0.69 [0.53, 0.85]	0.47 [0.22, 0.72]	0.03 [0.01, 0.05]
Keywords	0.21 [0.18, 0.24]	0.18 [0.12, 0.23]	0.51 [0.38, 0.63]	0.60 [0.45, 0.75]	0.49 [0.26, 0.71]	0.04 [0.01, 0.06]

Note. Values in brackets are 95% confidence intervals; D=Direct retrieval, R=Reconstruction, J_1=Familiarity-based judgment in the first test phase, J_{2r}, J_{2n}= Familiarity-based judgment in the first or second test phase given the judgment has been positive and negative, respectively, in the previous test phase; G=Guessing

Table 5. Estimated probabilities of verbatim and semantic imitation in T_1 in experiment 1

		Observed category		
		Tag reused (t_r)	Tag not reused (t_n)	Σ
Latent process	Verbatim imitation	0.20	0.00	0.20
	Semantic imitation	0.08	0.05	0.13
	No imitation	0.02	0.65	0.67
	Σ	0.30	0.70	1.00

Note. t_r=tag reused, t_n=tag not reused

Table 6. Contingency table separating the relative frequencies of reused tags by category ("verbatim response category" vs. "semantic response category") and output position (1 to 4)

Response category	Output position 1	2	3	4	Total
Verbatim	0.35	0.30	0.21	0.14	1.00
Semantic	0.23	0.24	0.25	0.28	1.00

Note. Verbatim response category = tags reused in each test phase, semantic response category = tags reused in only one or two test phase(s)

Table 7. Parameter estimates based on 2SDPg in experiment 2

Semantic Layout					
Type size	*Connectivity*	*D*	*R*	J_1	*G*
Large	*High*	0.09	0.10	0.52	0.03
	Low	0.08	0.10	0.45	0.03
Small	*High*	0.08	0.11	0.44	0.02
	Low	0.06	0.09	0.37	0.01
	M	**0.08**	**0.10**	**0.45**	**0.02**
Random Layout					
Type size	*Connectivity*	*D*	*R*	J_1	*G*
Large	*High*	0.09	0.18	0.51	0.02
	Low	0.05	0.19	0.62	0.00
Small	*High*	0.09	0.17	0.41	0.01
	Low	0.07	0.13	0.43	0.01
	M	**0.08**	**0.17**	**0.49**	**0.01**

Note. *D*=Direct retrieval, *R*=Reconstruction, J_1=Familiarity-based judgment in the first test phase, *G*=Guessing

Table 8. Estimated probabilities of verbatim and semantic imitation in T_1 as a function of tag cloud layout

	Semantic layout			*Random layout*		
	t_r	t_n	Σ	t_r	t_n	Σ
Verbatim imitation	0.08	0.00	0.08	0.08	0.00	0.08
Semantic imitation	0.04	0.05	0.09	0.08	0.08	0.16
No imitation	0.02	0.81	0.83	0.00	0.76	0.78
Σ	0.14	0.86	1.00	0.16	0.84	1.00

Note. t_r=tag reused, t_n=tag not reused; The estimates result from inserting the estimates of the parameter means of Table 7 into the model equations of Table 2.

Table 9. Mean LSA scores (and standard deviations) between newly assigned tags and tag clouds per artist by tag cloud layout

	Artist					
Group	Marina Black	Cartier-Bresson	Kruse-Jensen	Paul Leiter	Zoe Leonard	Helen Levitt
Random layout	0.120 (0.03)	0.191 (0.03)	0.163 (0.03)	0.171 (0.05)	0.175 (0.03)	0.163 (0.03)
Semantic layout	0.116 (0.03)	0.184 (0.04)	0.148 (0.03)	0.150 (0.04)	0.171 (0.03)	0.157 (0.04)

Appendix A

We can derive multiplicative terms from the processing tree of Figure 2. For instance, the equation (1) formalizes the processing tree's fourth path leading to the response category rnr,

$$P(\text{rnr}) = (1-D)RJ_1(1-J_{2\text{r}})J_{2\text{n}} \qquad \text{Equation (1).}$$

According to equation (1), the participant has not stored a verbatim but a gist trace of a particular tag in the study phase and therefore, reconstructs several semantically related words. Since the subsequent familiarity judgment selects the previously seen tag for output, $(1-D)RJ_1$, the participant's learning state in the test phase T_1 is P_r, a *partial recall* state underlying a tag's reuse. Since the STTT procedure does not include further study cycles, the participant will stay in the *partial recall* state for the last two test phases. However, the familiarity judgment can vary between T_1, T_2 and T_3 and the corresponding *partial* recall states can either be P_r or P_n that are associated with a positive and negative familiarity judgment, respectively. Equation (1) represents the case of a negative and positive familiarity judgment in the second and third test phase, respectively, associated with the learning states P_n in T_2 and P_r in T_3. The result of this sequence of cognitive events formalized by equation (1) is the probability of response category rnr. From the whole processing tree we can derive eight model equations (see Appendix B Table 10, column 2SDPg).

Appendix B

Table 10. Model equations of 2SDPg, 2SDP, 1SDPg and 2SOPg

	2SDPg	2SDP
rrr	$D+(1-D)[RJ_1J_{2r}^2+(1-R)G^3]$	$D+(1-D)[RJ_1J_{2r}^2]$
rrn	$(1-D)[RJ_1J_{2r}(1-J_{2r})+(1-R)G^2(1-G)]$	$(1-D)[RJ_1J_{2r}(1-J_{2r})]$
rnr	$(1-D)RJ_1(1-J_{2r})J_{2n}$	$(1-D)RJ_1(1-J_{2r})J_{2n}$
rnn	$(1-D)[RJ_1(1-J_{2r})(1-J_{2n})+(1-R)G(1-G)]$	$(1-D)[RJ_1(1-J_{2r})(1-J_{2n})]$
nrr	$(1-D)R(1-J_1)J_{2n}J_{2r}$	$(1-D)R(1-J_1)J_{2n}J_{2r}$
nrn	$(1-D)R(1-J_1)J_{2n}(1-J_{2r})$	$(1-D)R(1-J_1)J_{2n}(1-J_{2r})$
nnr	$(1-D)R(1-J_1)(1-J_{2n})J_{2r}$	$(1-D)R(1-J_1)(1-J_{2n})J_{2r}$
nnn	$(1-D)[R(1-J_1)(1-J_{2n})^2+(1-R)(1-G)]$	$(1-D)[R(1-J_1)(1-J_{2n})^2]$
	1SDPg	2SOPg
rrr	$G^3+(1-G)D+(1-G)(1-D)J_1J_{2r}$	$RJ_1J_{2r}^2+(1-R)G^3$
rrn	$G^2(1-G)+(1-G)(1-D)J_1J_{2r}(1-J_{2r})$	$RJ_1J_{2r}(1-J_{2r})+(1-R)G^2(1-G)$
rnr	$(1-G)(1-D)J_1(1-J_{2r})J_{2n}$	$RJ_1(1-J_{2r})(1-J_{2n})$
rnn	$(1-G)(1-D)J_1(1-J_{2r})(1-J_{2n})$	$RJ_1(1-J_{2r})J_{2n}+(1-R)G(1-G)$
nrr	$(1-G)(1-D)(1-J_1)J_{2n}J_{2r}$	$R(1-J_1)J_{2n}J_{2r}$
nrn	$(1-G)(1-D)(1-J_1)J_{2n}(1-J_{2r})$	$R(1-J_1)J_{2n}(1-J_{2r})$
nnr	$(1-G)(1-D)(1-J_1)(1-J_{2n})J_{2n}$	$R(1-J_1)(1-J_{2n})J_{2n}$
nnn	$(1-G)(1-D)(1-J_1)(1-J_{2n})^2$	$R(1-J_1)(1-J_{2n})^2+(1-R)(1-G)$

Note. D=Direct retrieval, R=Reconstruction, J_1=Familiarity-based judgment in the first test phase, J_{2r}, J_{2n}= Familiarity-based judgment in the first or second test phase given the judgment has been positive and negative in the previous test phase, respectively; G=Guessing

Table 11. Absolute frequencies of the STTT-response categories in experiment 1

	Tag	Keyword	Σ
rrr	269	272	454
rrn	20	20	40
rnr	13	16	25
nrr	20	23	49
rnn	36	43	65
nrn	9	15	18
nnr	15	20	38
nnn	770	743	1615

Table 12. Absolute frequencies of the STTT-response categories in experiment 2 for the semantic layout condition

	Semantic Layout						
	Lh	Ll	Sh	Sl	ΣL	ΣS	ΣΣ
rrr	136	114	120	78	250	198	448
rrn	14	12	9	8	26	17	43
rnr	11	8	13	11	19	24	43
nrr	23	19	25	24	42	49	91
rnn	35	35	30	18	70	48	118
nrn	10	10	11	17	20	28	48
nnr	14	15	18	14	29	32	61

nnn	909	939	926	982	1848	1908	3756

Note. Lh=large type size/high connectivity, Ll=large type size/low connectivity, Sh=Small type size/high connectivity, Sl=Small type size/low connectivity

Table 13. Absolute frequencies of the STTT-response categories in experiment 2 for the random layout condition

	Random Layout						
	Lh	Ll	Sh	Sl	ΣL	ΣS	ΣΣ
rrr	135	108	124	97	243	221	464
rrn	22	31	18	16	53	34	87
rnr	22	18	11	13	40	24	64
nrr	26	16	13	16	42	29	71
rnn	35	38	38	28	73	66	139
nrn	19	11	15	19	30	34	64
nnr	24	18	21	19	42	40	82
nnn	869	912	912	944	1781	1856	3637

Note. Lh=large type size/high connectivity, Ll=large type size/low connectivity, Sh=Small type size/high connectivity, Sl=Small type size/low connectivity

Table 14. Parameter estimates based on 2SDPg in experiment 2

Semantic Layout							
Type size	*Connec-tivity*	D	R	J_1	J_{2r}	J_{2n}	G
Large	*High*	0.09 [0.06, 0.12]	0.10 [0.07, 0.15]	0.52 [0.36, 0.68]	0.70 [0.54, 0.85]	0.58 [0.34, 0.81]	0.03 [0.01, 0.05]
	Low	0.08 [0.06, 0.11]	0.10 [0.06, 0.15]	0.45 [0.28, 0.62]	0.66 [0.49, 0.82]	0.48 [0.20, 0.75]	0.03 [0.01, 0.04]
Small	*High*	0.08 [0.06, 0.11]	0.11 [0.08, 0.15]	0.44 [0.29, 0.59]	0.67 [0.51, 0.82]	0.58 [0.38, 0.78]	0.02 [0.01, 0.04]
	Low	0.06 [0.04, 0.07]	0.09 [0.07, 0.11]	0.37 [0.23, 0.50]	0.58 [0.43, 0.72]	0.67 [0.48, 0.86]	0.01 [0.0, 0.02]
Random Layout							
Type size	*Connec-tivity*	D	R	J_1	J_{2r}	J_{2n}	G
Large	*High*	0.09 [0.06, 0.12]	0.18 [0.13, 0.23]	0.51 [0.40, 0.62]	0.56 [0.43, 0.70]	0.51 [0.32, 0.67]	0.02 [-0.0, 0.04]
	Low	0.05 [0.02, 0.09]	0.19 [0.13, 0.25]	0.62 [0.48, 0.75]	0.59 [0.41, 0.77]	0.34 [0.15, 0.54]	0.00 [-0.0, 0.03]
Small	*High*	0.09 [0.07, 0.12]	0.17 [0.08, 0.26]	0.41 [0.27, 0.56]	0.46 [0.28, 0.64]	0.28 [0.07, 0.48]	0.01 [-0.02, 0.04]
	Low	0.07 [0.05, 0.09]	0.13 [0.08, 0.19]	0.43 [0.31, 0.55]	0.47 [0.31, 0.62]	0.42 [0.17, 0.67]	0.01 [-0.01, 0.03]

Note. Values in brackets are 95% confidence intervals; *D*=Direct retrieval, *R*=Reconstruction, J_1=Familiarity-based judgment in the first test phase, J_{2r}, J_{2n}= Familiarity-based judgment in the first or second test phase given the judgment has been positive and negative in the previous test phase, respectively; *G*=Guessing