

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

ANGELINA DE CARVALHO ALVAREZ ZIESEMER

**A LANGUAGE-BASED APPROACH TO SUPPORT THE IDENTIFICATION OF TAGGING
BEHAVIOUR**

Porto Alegre

2017

PÓS-GRADUAÇÃO - *STRICTO SENSU*



Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
FACULTY OF INFORMATICS
COMPUTER SCIENCE GRADUATE PROGRAM

**A LANGUAGE-BASED APPROACH TO SUPPORT THE
IDENTIFICATION OF TAGGING BEHAVIOUR**

ANGELINA DE CARVALHO A. ZIESEMER

Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Ph.D. in Computer Science.

Advisor: Prof. Dr. Milene Selbach Silveira

**Porto Alegre
December, 2017**

Ficha Catalográfica

Z67L Ziesemer, Angelina de Carvalho Alvarez

A Language-based Approach to Support the Identification of Tagging Behaviour / Angelina de Carvalho Alvarez Ziesemer . – 2017.

133.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Milene Selbach Silveira.

1. Tagging. 2. Human-Computer Interaction. 3. Semiotic. 4. Recommender Systems. I. Silveira, Milene Selbach. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Salete Maria Sartori CRB-10/1363

Angelina de Carvalho Alvarez Zieseimer

**A Language-based Approach to Support the
Identification of Tagging Behaviour**

This Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on December 14th, 2017.

COMMITTEE MEMBERS:

Prof. Dr. Rodrigo Coelho Barros (PUCRS)

Prof. Dr. Raquel Oliveira Prates (UFMG)

Prof. Dr. Lucia Vilela Leite Filgueiras (USP)

Prof. Dr. Milene Selbach Silveira (PUCRS)

ACKNOWLEDGEMENT

First I would like to thank my husband and friend Adriel Ziesemer Junior for being supportive and encourage me to reach my goals. To my son Daniel: you bring light to our lives, I am better because of you. I would like to sincerely thank my thesis adviser, Milene Selbach Silveira, for all her support, guidance and friendship throughout this time. Thanks for the following professors: João Batista de Oliveira, Isabel Mansour, James Blustein, thanks for the encouragement and support. Thanks to my friends and colleagues, Anderson Silva, Ricardo Piccoli, Andriele Busatto, Rodrigo Chamun, Anna Cunha, and Daniele Souza. I learned a lot of things with you guys and enjoyed the happy moments as well. Special thanks to my friends Luana Muller and Vinicius Cassol, we shared this journey together, we made it! And finally and most important, thank you God, for giving me the strength to never give up.

A LANGUAGE-BASED APPROACH TO SUPPORT THE IDENTIFICATION OF TAGGING BEHAVIOUR

ABSTRACT

Tags work as a tool to support search engines on the task of finding resources by their subject matter and the content they bring. However, users have expanded tag functionality with the intent of expressing opinion, personal categorization, and even as a tool for spamming propagation. Users can provide a great deal of qualitative data about their motivations for tagging, but little has been explored regarding tagging behaviour and patterns from a quantitative and contextual point of view. As tags are basically keywords that users resort as a tool for describing content, we analyzed their use from a linguistics perspective. The results we found during a set of user studies we conducted supported us on the task of designing a language-based approach that rely on tagging patterns as quantitative data for the identification of tagging behaviour. During a case study to analyze our approach, we used real datasets to compute the features we defined in our model in combination with clustering tools applied to datasets from Flickr and Instagram, resulting in personas that explain users' motivation for tagging. We were able to point to the differences among tagging behaviour and how the choice of structure or language for tagging could be used as source to identify users' motivation for tagging when sharing content online. We found that the patterns and motivation we have modeled in our approach replicate in real datasets. This could benefit those who wish to use tags as source for a variety of projects, such as modeling users' behaviour through tags available online, choosing a recommendation approach based on users' motivations for tagging, preselecting data and tags as source for recommendation according to system goals and user needs, identifying users opened to receive contextual content, among others.

Keywords: Tagging, Behaviour, Hashtags, Semiotic.

UMA ABORDAGEM BASEADA EM LINGUAGEM PARA APOIAR A IDENTIFICAÇÃO DE PADRÕES DE COMPORTAMENTO NO USO DE TAGS

RESUMO

Tags são ferramentas que apoiam engines de busca na tarefa de encontrar conteúdo relacionado a assuntos que estes contém. Entretanto, usuários de tags online expandiram a funcionalidade das tags com o objetivo de expressar opinião, categorização pessoal de conteúdo e até mesmo para propagação de spams. Quando se analisa um conjunto de tags, pode-se perceber que tags são fontes de uma quantidade significativa de dados qualitativos que estão relacionados à motivação do seu uso. Entretanto, pouco tem sido investigado em relação a padrões de uso das mesmas de um ponto de vista quantitativo e contextual. Tags são basicamente palavras-chave que usuários online usam como uma ferramenta para descrever conteúdos e portanto nós analisamos o seu uso de um ponto de vista linguístico. Os resultados que encontramos durante um conjunto de estudos com usuários serviu de apoio ao desenvolvimento de uma abordagem linguística que conta com padrões de tags como dado quantitativo para identificação de comportamento de utilização das mesmas. Durante um estudo de caso para analisar a abordagem proposta, nós utilizamos datasets de usuários de tags online (Flickr e Instagram) para calcular as características definidas no nosso modelo em combinação com ferramentas de agrupamento, o qual resultou na modelagem de Personas para explicar a motivação dos usuários quando utilizam tags. Nosso trabalho foi capaz de identificar diferenças no comportamento de utilização das tags e como que a escolha da estrutura e idioma escolhido para as mesmas pode servir como fonte para identificação da motivação para o seu uso quando compartilhando conteúdo online. Também foi possível identificar a replicação desses padrões e motivações que modelamos nos datasets coletados em redes sociais. Nós acreditamos que esta abordagem pode beneficiar aqueles que tem acesso a tags como fonte para modelagem de usuários, tais como, na escolha de tipo de abordagem de recomendação com base na motivação do usuário, na pré seleção de dados para recomendação de acordo com a necessidade dos usuários e do sistema, na identificação da abertura do usuário em relação a recebimento de conteúdo contextual online, entre outros.

Keywords: Tags, Comportamento, Usuário, Hashtags, Semiótica.

LIST OF FIGURES

Figura 2.1	Multilanguage tags and a sentence used to classify the image. These are tags from Flickr before the new rules for adding tags started in 2012. Flickr used to allow sentences as tags to represent a photo.	30
Figura 2.2	Flickr used to allow composed tags, but currently the space among words represents that each word is a different tag.	30
Figura 2.3	Tags assigned by a robot.	31
Figura 2.4	Differences among datasets, designers, and users.	37
Figura 2.5	The dimensions of syntagmatic and paradigmatic relations.	38
Figura 3.1	Example of types of images used in this study.	43
Figura 3.2	P_a classified as an image with fair context representation for its content. . .	44
Figura 3.3	P_b classified as an image with high context representation for its content. . .	45
Figura 3.4	P_c classified as an image with high context representation for its content. . .	45
Figura 3.5	P_d classified as an image with fair context representation for its content. . .	46
Figura 3.6	P_e classified as an image with fair context representation for its content regarding the image situation. It is important to point that this was an image that brings location as context information as well.	46
Figura 3.7	P_f classified as an image with fair context representation for its content regarding the image situation. It is important to point that this was an image that brings location as context information as well.	47
Figura 3.8	P_g classified as an image with fair context representation for its content. . .	47
Figura 3.9	Power-law distribution comparing NR vs. RS stage (A) and the distribution of reference and recommended tags (B) from Brazil.	50
Figura 3.10	Power-law distribution comparing NR vs. RS stage (A) and the distribution of reference and recommended tags (B) from Canada.	51
Figura 3.11	Comparing participants' language chosen for tagging in both NR and RS stages.	55
Figura 3.12	Languages adopted on photos presented in both stages.	55
Figura 5.1	Indexing and Contextualization, motivation dimensions versus structure chosen for tagging.	68
Figura 5.2	Tendency among Portuguese speakers: assigning putative tags (paradigmatic structure, in general), popular tags (memes, pool of images related to tags – #ThrowbackThursday), or accepting tag recommendation.	70
Figura 5.3	All tags go to the first step of language identification through a library called Enchant. The tag #bestFriends went through the process of Tag segmentation since it was not identified as a word present in Portuguese and English dictionaries.	72

Figura 5.4	Language identification steps during the identification process of paradigmatic tags.	73
Figura 5.5	Example of a user that has three posting P . Each posting is composed by an image (r_i), and set of tags T in which $T = t_1, t_2, \dots, t_n$	74
Figura 5.6	Word segmentation approach used for the identification of tags that have syntagmatic structure.	75
Figura 5.7	Even when words have no real meaning, the algorithm will try to split them.	76
Figura 5.8	Difference in use of repeated tags.	78
Figura 5.9	Heterogeneity of tags and the difference in the use of tags as resource for content indexing.	78
Figura 6.1	Cut off of users with less than 20 tags assigned in total – Instagram dataset, English.	85
Figura 6.2	Cut off of users with less than 20 tags assigned in total – Instagram dataset, Portuguese.	86
Figura 6.3	Flickr dataset and its data cut off to reduce sparsity and improve the quality of results in the clustering task.	87
Figura 6.4	Comparison between two approaches for clustering data: K-means vs. EM.	90
Figura 6.5	Comparing two approaches for clustering data. K-means, vs. EM.	91
Figura 6.6	Differences in mean levels of feature <i>paradigmatic tags assigned in English (ENG_para)</i> . This feature is compared pair by pair for each of the nine fitted clusters.	94
Figura 6.7	Same approach applied to evaluate the clustering performance of a different number of clusters. Fitting five clusters, the results show that the mean is not the same when comparing clusters pair by pair.	95
Figura A.1	Letter of approval for conducting research with humans at PUCRS – Brazil.	126
Figura A.2	Letter of approval for conducting research with humans at Dalhousie University – Canada.	127

LIST OF TABLES

Tabela 3.1	Results from content-oriented classification of photos.	44
Tabela 3.2	Classification of tag structures from participants from Brazil for the photos presented in both stage for each group. The p -value ($p < 0.01$) shows that there is an association between the type of system used and the type of tag structure assigned.	47
Tabela 3.3	Classification of tag structures from participants from Canada for the photos presented in both stage for each group. The p -value ($p < 0.01$) shows that there is an association between the type of system used and the type of tag structure assigned.	48
Tabela 3.4	Proportion of syntagmatic and paradigmatic tags in both stages of the experiment to each image assigned by participants from Brazil.	49
Tabela 3.5	Proportion of syntagmatic and paradigmatic tags in both stages of the experiment to each photo (Canada).	49
Tabela 3.6	The list of most frequently assigned tags, their class and the representative proportion among the set of tags of the same photo*.	53
Tabela 3.7	Comparing the proportion of images assigned mainly in PT and EN in both stage of the experiments.	54
Tabela 3.8	The list of most frequently assigned tags, their class and the representative proportion among the set of tags of the same photo*.	56
Tabela 5.1	General motivations defined based on the subset of motivations already describe by the literature [GLYH10].	66
Tabela 5.2	Tagging structure versus the motivations for tagging.	67
Tabela 5.3	Indexing and Language Domain were two main factors found to lead users to assign or not tags in foreign languages.	70
Tabela 5.4	Framework features to support the identification of tagging behaviour. . . .	80
Tabela 5.5	Overview of the features as a final result computed by the framework. . . .	81
Tabela 6.1	Instagram dataset and summary of framework results.	85
Tabela 6.2	Instagram dataset resulted after the cut off of <i>ids</i> that had less than 20 assigned tags. Our goal is to decrease data sparsity and improve results of the clustering analysis we will conduct.	87
Tabela 6.3	Flickr dataset, before and after data cut off.	88
Tabela 6.4	P -value for a pair of clusters. Comparison regarding the target feature we analyzed: <i>paradigmatic tags assigned in English (ENG_para)</i> for nine clusters.	93
Tabela 6.5	Multi comparison of features for five clusters. The p -values show that there is difference among the means resulting for the target feature being analyzed.	94

Tabela 6.6 Five clusters fitted by EM based on a Gaussian Mixture Model for the dataset of English speakers. 96

Tabela 6.7 Five clusters fitted by EM based on a Gaussian Mixture Model for the dataset of Portuguese speakers. 97

Tabela 6.8 Flickr clustering results for both Portuguese and English language. 99

Tabela 6.9 Personas based on clustering results from Instagram datasets. 104

Tabela 6.10 Personas based on clustering results from Instagram datasets. 105

Tabela 6.11 Personas based on clustering results from Instagram datasets. 106

Tabela 6.12 Personas for Flickr dataset we have analyzed. 107

Tabela 6.13 Flickr Personas 108

Tabela C.1 URL locations for the images used during the user studies. 133

LIST OF ACRONYMS

API	<i>Application Program Interface</i>
URL	<i>Uniform Resource Locator</i>
NLP	<i>Natural Language Processing</i>
GPL	<i>General Public License</i>

CONTENTS

1. INTRODUCTION	23
1.1 Motivation	24
1.2 Scope of Research	24
1.3 Goals	25
1.3.1 Specific Goals	26
1.4 Method and Instrument Design	26
1.5 Thesis Proposal Organization	27
2. THEORETICAL BACKGROUND AND RELATED WORKS	29
2.1 Tagging	29
2.1.1 Tag types	31
2.1.2 Tagging Behaviour and Motivation	33
2.2 Recommender Systems	34
2.2.1 Tagging Recommendation	35
2.3 Semiotic	37
3. USER STUDY	41
3.1 Method and Instrument Design	41
3.1.1 Tagging using Recommendation	42
3.1.2 Content-based classification of Images	43
3.1.3 Processing of Tagging Dimensions	44
3.1.4 Programmatically Classifying Languages	45
3.2 Findings	46
3.2.1 Tagging Structure	46
3.2.2 Language	51
3.3 Discussion	57
4. UNDERSTANDING USERS' TAGGING PATTERNS	59
4.1 Method and Instrument Design	59
4.2 Results	60
4.2.1 Language Choice	60
4.2.2 Structure Choice	62
4.3 Discussion	63

5. A LANGUAGE-BASED APPROACH TO SUPPORT THE IDENTIFICATION OF TAGGING BEHAVIOUR	65
5.1 A Model of Tagging Patterns and Its Dimensions	65
5.1.1 Structure	66
5.1.2 Language	69
5.2 A Framework for the Identification of Tagging Behaviour	70
5.2.1 Language Identification	72
5.2.2 Tag Segmentation	74
5.2.3 Repetition and Heterogeneity of tags	77
6. CASE STUDY	83
6.1 Data Collection, Data Pre-processing, and Framework Application	83
6.1.1 Instagram	83
6.1.2 Flickr	85
6.2 Clustering	88
6.2.1 Clustering Tools	89
6.2.2 Multiple Comparisons of Features	91
6.3 Findings	92
6.3.1 Instagram Clusters	93
6.3.2 Flickr Clusters	97
6.4 Tagging-Based Personas	98
6.4.1 Personas for Instagram dataset	100
6.4.2 Personas for Flickr dataset	102
7. FINAL CONSIDERATIONS	109
7.1 Limitations	112
7.2 Future Works	112
8. Activities During Period as Ph.D. Student	115
Bibliography	117
A. Appendix A	125
A.1 PUCRS Approval Letter	125
A.2 Dalhousie University Approval Letter	125

B. Appendix B	129
B.1 Consent Form Dalhousie University	129
B.2 Consent Form PUCRS	130
C. Appendix C	133
C.1 Images	133

1. INTRODUCTION

Nowadays, global systems, social networks, and the like, must deal with a wide range of user-generated data that involve a combination of social factors, such as cultural differences, age, gender, language, among other characteristics and motivations that model individual behaviour. Tags, or hashtags, have become a powerful tool for indexing content on social networks, websites and systems. Although tags have been used with the purpose of recommending content, inferring user similarities, finding out about events, as source for the identification of users' opinion [QSCT⁺17], and even aiming at recommending new tags during the tagging task [QCDM⁺11, ZO11, SNPAR14], they were primarily introduced with the purpose of allowing users to classify image content. Tags work as a tool to support search engines on the task of finding resources by their subject matter and the content they bring. However, users have expanded tags' functionality with the intent of expressing opinion, personal categorization, or even as a tool for spamming propagation. As a result, this generates several types of tags that are associated with different motivations for tagging [SS16, AN07, GLYH10].

As the amount of online data increases, modeling users' behaviour provides valuable information to answer questions regarding users' motivation, preferences, and goals when using an application. This is especially important for applications that rely on social media APIs and/or user-generated data as source for propagation of content. According to Turkley [Tur05], when different people sit down in front of computers, even when they are supposed to execute the "same" task, their interacting styles can be very different, generating unexpected outcomes and, consequently, differing users by their behaviour while using a tool, system, or application. As designers, researchers, and developers, we should take into consideration that users are different, and applications that rely on "one-size-fits-all" approaches could fail to attend users' needs [KRW11].

In April 2017, Instagram reached the total of 700 million active users per month, and 1 million active advertisers [Con17]. Instagram number of users is twice as higher as the number of users on Twitter, which shows how fast this photo sharing application has grown. The analysis of tagging behaviour can suit as source to identify users' profile and their intentions when tagging or even their motivations when using a system. As a future benefit, this could support designers on the choice of data/users/patterns of tags to be used as source for tag recommendation algorithms, to identify users according to their motivations for using a system, and so on. However, tag patterns are mostly investigated through qualitative data analysis, which demands knowledge of the tagging field [AN07, EG12, GKE11] and expertise on user experience research. This is not suitable for designers, researchers, and project managers, who often face problems of low or no budget to conduct user research, which is time-consuming and demands specific knowledge in this field [Use17]. In fact, research has reported that it is common for designers/developers/engineers to create assumptions of users' preferences based on their own experience [Nie12]. In addition to that, it has become an impossible task to evaluate tagging behaviour using manual approaches exclusively, due to the amount of data generated online.

In this work we present an approach to support the identification of tagging behaviour through tagging patterns as an additional source for modeling users' behaviour. This could benefit those who intend to use tags as source for a variety of projects, such as modeling users' behaviour through tags available online, choosing a recommendation approach based on users' motivations for tagging, preselecting data and tags as source for recommendation according to system goals and user needs, identifying users opened to receive contextual content, among others.

1.1 Motivation

Today many systems and applications are resorting to users' data available online¹ – providing designers with the possibility to integrate systems and applications with social media log in, and providing easy access for users to connect to new services and networks. This results in a huge variety of users from different backgrounds, cultures, and languages, making use of the same application or platform. When designers face such situation, conducting user research essentially through qualitative analysis may not be enough, that is, the target population available online is too broad to conduct research using tools exclusively designed for qualitative analysis due to the amount of data available [Max08]. Users can provide plenty of qualitative information about their motivations for tagging, but little has been explored regarding tagging behaviour and patterns from a quantitative point of view [KKGS10,SKK12]. Moreover, there are many references that support the identification of types of tags based on users' motivation for tagging [DF10, AN07, SLR⁺06, GKE11], yet, in a general way, this approach can be essentially carried out through manual analysis that depends on experts' insight.

Recently, Facebook corporation has reported their concern about the way users are sharing content online. They named it “the context collapse”, regarding the lack of personal information users are sharing [Gri16] and their tendency to move to more personal services, such as Instagram, in which short content descriptions are common and tags are highly used. Instead of trying to identify users' preferences and motivation for using a system based on the data provided on their public profile, designers can resort to other sources of data publicly shared online to support the process of conducting user research.

The combination of data extracted from real datasets, data analysis, and human insight could support designers in the task of identifying users' behaviour as well as users' profile, preferences, and goals. Based on the research we conducted in this work, we believe that it is possible to identify tagging behaviour, and users' motivation for tagging through a language-based approach.

1.2 Scope of Research

Since tags are basically keywords which users resort to as a tool for describing content, we analyzed their use from a linguistic perspective [DS11]. We intend to be able to identify groups

¹APIs from Facebook, Flickr, Instagram, Quora, and Twitter allow developers to use data publicly available online and tools for integrating user personal log into applications.

of users that are subject to tagging behaviour identification through a combination of features related to patterns of tagging. In order to address this task, research conducted in this thesis is focused on the study of users' choices and patterns for tagging, especially regarding its structure, language, and how it is related to users' motivation. This research has started with general goals that were refined as we started having insight about how users' tagging behaviour could be related to motivations for tagging. We started our research with a small scope that led us to answer a more comprehensive question. First, we conducted some user studies based on findings about a previous work regarding tagging recommendation [ZO11]. Based on the results found at that stage, we hypothesized that users do not differ in the structure and language used for tagging. In order to investigate this hypothesis, we performed a combination of user studies with participants from two different countries (Canada and Brazil). As the studies evolved and the hypotheses were analyzed, we narrowed our investigation down in order to answer the following research questions:

1. Are there any patterns of tags (structure and language) that can contribute to the identification of tagging behaviour?
2. How are patterns of tags, regarding structure and language, related to users' motivation for tagging?
3. Is it possible to automatically identify tagging patterns to support the identification of tagging behaviour?

We conducted user studies, that helped us identify differences in the language and structure used for tagging. This led us to model tagging behaviour through the combination of tagging patterns and users' motivation for tagging. The results we found assisted the task of designing a language-based approach that rely on tagging patterns as quantitative data for the identification of tagging behaviour. During a case study, we used real datasets to compute the features we defined in our approach in combination to clustering tools applied to datasets from Flickr and Instagram, resulting in personas that explain users' motivation for tagging.

1.3 Goals

This work aims to support rather than replace any type of user modeling approach. We try to identify the intentions behind assigned tags, so that one can use such information as source for designing any system or application that can rely on this type of data. The findings achieved in this work supported the design of a tagging pattern model and a framework to support the identification of tagging behaviour.

As a general goal, we aim to provide, through our model, insights about users' tagging behaviour and motivation for tagging. Consequently, through the application of our approach, designers shall be able to identify tagging behaviour and create more inclusive applications that use diverse data, based not only on generalization, but on group segmentation that includes distinct users who

may be underrepresented. Results found in this work can support the choices of data sources for recommender systems or any other application for developing approaches that could rely on tagging sources.

1.3.1 Specific Goals

The specific goals we aim to reach at the end of this research are presented as follows:

- Understanding the aspects of tagging behaviour, not only tagging patterns;
- Identifying the differences between Portuguese and English speakers on their choices for tagging.
- Understanding how motivation is related to the choices of tagging patterns;
- Modeling tagging behaviour;
- Designing an approach to support the identification of tag patterns and the influence of repetition and variability on these patterns.

1.4 Method and Instrument Design

In this work we used mixed methods of research that consist in a combination of tools to gather and analyze qualitative and quantitative data.

The hypotheses we present in this work aroused in the beginning of our investigation and were based on the results of a previous work [dCZdO13], developed by the author of this thesis. In order to answer the research questions we have proposed, we conducted an investigation that consists of three stages. The first stage of this work consists of the investigation of related works and background in the field of tagging behaviour later related to our approach. Secondly, we conducted a user study using a within-subjects manipulation [LFH17] of tagging with system recommendation/non-recommendation (counterbalanced order), in which participants from Brazil and Canada (comparison between-group) were invited to join. The third stage of the user study consists of the use of a survey (mixed close-ended and open-ended questions) and interviews (semi-structured open-ended questions), so that an in-depth understanding of motivations for tagging could be achieved.

The data gathered during the studies were analyzed using statistical analysis when quantitative data were involved, and textual analysis and Grounded Theory tools when data were gathered from open-ended questions.

To validate the approach, based on the cited studies, we resorted to a case study using real tagging datasets. We analyzed the approach outcomes through a combination of a clustering algorithm and the use of Personas [Nie12].

1.5 Thesis Proposal Organization

The remainder of this thesis is organized as follows. We begin by presenting the theoretical aspects later related to our approach and findings, and the related works we investigated in the field of tagging behaviour analysis in the next chapter. After that, we present the user studies we have conducted with participants from Canada and Brazil, followed by the explanation of the research done on the obtained data, as well as our findings. Chapter 5 presents the approach we have created to support the identification of tagging patterns and a framework for the task of identifying tagging behaviour. We conducted a case study in which we resorted to the support of clustering tools in order to identify groups of users on datasets from Flickr and Instagram. Moreover, we resorted to the use of Personas to explain the clusters resulted from the algorithm used. Finally we present the discussion on the results and the contribution of this work.

2. THEORETICAL BACKGROUND AND RELATED WORKS

In this section we present the theories and techniques that served as foundation to perform the experiments and to design the language approach we propose in this work. In addition, we present related works in the field of tagging behaviour that will be later related to our findings for the language approach we designed.

2.1 Tagging

Imagine you are publishing a collection of photos on Flickr, Instagram or other social media network. If you want your photos to be found by other users or even by yourself while browsing, you will probably add tags to your photo collection as a way of indexing it. User-generated tags for classifying content are known as folksonomy (folk + taxonomy), and they are generally performed by regular users on the web. It has been a common approach used by regular users for content classification and it was created with the initial purpose of allowing users to describe images, videos, text, or any other type of personal content or resource by the attribution of “words” to resources [GLYH10]. Tags also allow user browsing resources and search engines to obtain content through keywords and queries. More specifically, social network services allow users to assign hashtags, with the purpose of helping them in the navigation of content by subject. Tags on social media networks are still one of the main resources to assist the retrieval of photos and videos, the types of resources that does not bring any textual description regarding its content.

Many applications and social networks, such as Facebook, Instagram, Twitter, Tumblr, Youtube, among others, resort to the use of tags as a tool for indexing content. On Twitter, a microblogging service, tags are used especially as a conversational and organization tool [HTE10]. It is common for users to adopt tags into sentences, promoting a way to categorize content and introduce a subject matter at the same time (e.g. “*Can’t wait to watch the #oscar tonight!*”), along with the possibility to follow a discussion from the same topic. Also, Twitter tags play an important role as source for topic search and for the discovery of new content based on trending topics that are indexed as hashtags [Far17].

On Instagram, tags are used as a tool to raise interactions. Since the *like* resource was introduced as an indicator of popularity on the network, users try to collect more “likes” and followers through the tags they use, for instance, *#like4likes*, *#followfriday*, *#follow4follow*, *#tagsforlikes* [ZNHP17]. In the work of Araújo et al. [ACdS⁺14], they found evidence that the popularity of a tag can be driven by the number of followers a user has, and that tags attract people who are not in a target user-follower list but are interested in a specific topic (e.g. *#rockinrio*).

Regarding the use of tags on Flickr, we have noticed in the course of time the difference in the way its interface allows interaction and tag management. Flickr used¹ to allow users to add

¹The image was first posted in February, 2009 and tags were assigned at the same time.



Figure 2.1: Multilanguage tags and a sentence used to classify the image. These are tags from Flickr before the new rules for adding tags started in 2012. Flickr used to allow sentences as tags to represent a photo.

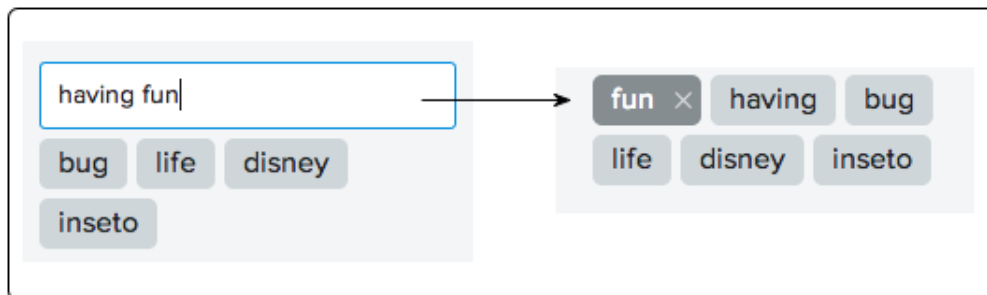


Figure 2.2: Flickr used to allow composed tags, but currently the space among words represents that each word is a different tag.

sentences as tags (composed tags, with space among words, as in “*vida simples*”²) as we can see in Figure 2.1. Nowadays³ whenever a tag is added with a space on it, such as *having fun*, the system will automatically identify it as two separated keywords (Figure 2.2). It also has an approach for the identification of duplicated tags and it resorts to a robot which will automatically suggest tags to describe the image content, as we can see in Figure 2.3.

²translation from Portuguese to English – “simple life”

³Current version accessed in October, 2017.

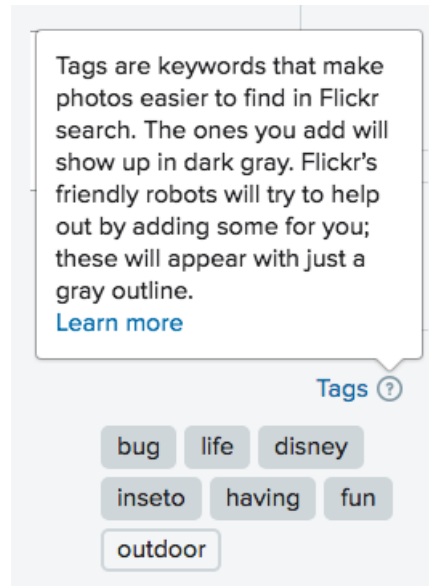


Figure 2.3: Tags assigned by a robot.

Regarding the way users assign tags, Figure 2.1 shows a photo from a Flickr user that uses sentences, single words and two languages (Portuguese and English) for tagging. Multilanguage tags [SGP11, ZBS16b] are a behaviour users adopt for tagging, and, despite their advantages, it is common to find tags with typos, as we can see in the same Figure 2.1, where the tag “*balon*”[sic] was used instead of “*balloon*”. Thus, the acceptance of suggested tags by recommender systems can promote a common vocabulary among the social media community and improve the results of tag-based searches.

It is important to point out that in many cases the differences in tagging practice can be caused by the design and function of an application or system [HTE10]. Research has found that some users do not assign any tags to their content simply because they do not know what types of tag they should use [SLR⁺06]. The recommendation of tags has been investigated as a tool for promoting the use of tags and increasing tags reuse on social tagging systems [dCZdO13, KPL17]. However, investigations in the field of tagging behaviour have shown that the tag choices are influenced by the recommending approaches used for supporting the tagging task [ZBS16b, SLR⁺06]. This topic will be discussed further on as we present the theoretical considerations that can lead users to change behaviour as they use recommender systems and the reasons we decided to use these tools to identify users’ preferences for tagging.

2.1.1 Tag types

In the work of Gupta et al. [GLYH10], the authors presented a systematic study on tagging literature. They provided a detailed classification of users’ motivation for tagging and the types of tags used. According to them, users are mainly motivated by:

- Future retrieval (helpful tags for search engines and for later browsing);

- Description, sharing and contribution (tags which can provide cluster regarding a subject or categories);
- To call attention;
- Play and competition (users can add a tag from an organization to compete for prizes);
- Self-presentation (the name of the content owner, or a tag like “mystuff”);
- Opinion expression (“good”, “worst”, “likeit”);
- Task organization (“toread”, “todo”);
- Contextual information (to communicate content context to others);
- To earn money (organizations paying users to assign a tag).

After classifying users' motivation, they categorized the types of tags users assign:

- Content-based tags (tags that describe content);
- Context-based tags (location, time);
- Attribute tags (who the resource refers to);
- Ownership tags (resource owner);
- Subjective tags (opinions, emotions);
- Organizational tags (“toread”, “todo”);
- Purpose tags (not related to content);
- Factual tags (facts about people, places and concepts);
- Personal tags (self-reference and content organization);
- Self-referential tags (resources that refer to themselves);
- Tag bundles (hierarchical folksonomies, as an URL to another URL).

Veres [Ver06] presented a linguistic classification of tags. This classification was created based on the function of assigned tags: functional tags (describing the function of the resource being tagged), functional collocation, origin collocation (describing the reason things are together through tags such as “trip”, “vacations”, “family”), function and origin (why and where they come from), taxonomic (words used for classification), adjectives (funny, great), verbs, and proper names.

2.1.2 Tagging Behaviour and Motivation

Researchers have investigated how and what types of tags users assign, and their motivation to do so [AN07, KKGS10, EG12, GKE11, SLR⁺06, SGP11]. Because tagging behaviour can be covered and analyzed from several aspects, the research conclusions we found resulted in a variation of distinct concepts that cover tags as classes, functions, types, vocabulary, and motivations.

Eleta and Golbeck [EG12] investigated social tagging patterns in English and Spanish, and found out that the level of agreement (the same word meaning assigned from different language speakers; e.g. “*dog*”, “*perro*”⁴) of tags in both languages to describe images does not change significantly. Also, Stiller et al. [SGP11] found indicators that the resource language does not correlate with the language used for tagging. This last study was specific on tags assigned to articles shared online through BibSonomy⁵. Researchers collected articles that had mainly German content and found that most of them were tagged using both English and German. However, their research still needs a deeper investigation on the subject, since 86% of the analyzed URLs were only tagged by one user.

Regarding motivation for tagging, as far as we know, the study conducted by Morgan and Naaman [AN07] was the first one to introduce social aspects as motivation for tagging. They created a taxonomy based on social motivation on ZoneTag and Flickr under a qualitative study that analyzed data from 13 participants who took part in a semi-structured interview. Motivation for tagging was classified according to the target “audience” and tag “function”. Users motivations for tagging appeared to be related to the audience the author intended to send the tags to, which can include – the author him/herself, friends and family, and the general public.

In the work of Sen et al. [SLR⁺06], they manually classified 3,263 distinct tags into general classes (Factual, Subjective, and Personal) [GH06] in order to relate them to the MovieLens community usage of tags. They concluded that personal tendency, past tagging behaviours and the influence of other users in a community can affect users’ choices for tagging.

Regarding tagging and culture, Dong [DF10] found that there are cultural differences on tagging behavior between European Americans and Chinese. They performed an experiment with 44 participants (21 European Americans and 23 Chinese). Sixty digital photos from Google⁶ were presented to participants, each one having real-life objects, no language or cultural icons and at least a clear foreground main object as well as a variety of background objects. Besides that, they asked participants to use single words to describe the images rather than long phrases or sentences. Researchers manually coded the tags used by users into the following categories: foreground main object, background object, overall description and relationship (tags that describe the relationship among objects on the image). Description of foreground photos occurs earlier (related to tag insertion order) on tags added by European Americans, while Chinese are more likely to describe foreground and background equally (regarding tag insertion order). Also, still regarding tag position, European Americans add

⁴Perro is the translation of “dog” from Spanish to English.

⁵BibSonomy is an online social bookmarking and publication-sharing system

⁶<http://google.com>

tags referring the object name more often than Chinese, while tags describing the overall image content tend to be added by Chinese in their first tags.

Regarding tagging behaviour, we found that most of the researches conducted in this field use manual coding for the identification of tagging classes, types, or patterns [AN07, SLR⁺06, DF10, GKE11]. Korner et al. [KKG10], conducted a research overview on users' motivation for tagging and the type of detection used for its identification. They presented 11 relevant pieces of research in the field, from which 10⁷ of them used experts' judgment to classify users' behaviour or motivation for tagging. The exception of such collection was their own work [SKK10], which has used primary quantitative analysis to classify users between categorizers and describers. Their research had an important contribution for one of the steps we use for the identification of tagging behaviour. Based on the users' motivation for tagging, they found that the reuse of tags is a dimension that could be related to categorizers, since these users' main goal is browse their content at a later time. We did not use their same measures due to the nature of our model, but, due to their work, we are aware that tag repetition may indicate categorization of content.

Although many motivations and types of tags have been identified in the previously presented studies, we did not find research pointing to the relation among tag function/motivation and the structure and language assigned during the tagging task.

Next, in Section 2.2, we present the tools used to conduct the user study we shall present in Chapter 3 and the language theory used (Section 2.3) as basis for the creation of our approach.

2.2 Recommender Systems

Due to the advent of the Web 2.0, users are willing to contribute to their community knowledge by tagging, adding photos, videos and text on social media/social networks, and this behaviour has increased the number of content available online. The task of finding relevant resources through the web could be a disappointing task. The reason for that is that among a lot of available content online, users need to filter a huge list of searching results to find what they are looking for. The purpose of recommendation algorithms is to filter resources according to users' preferences and there are distinct approaches [BOH12, RAC⁺02, BS97] developed to try to accomplish such task. However, relying on recommendation and predictive algorithms without understanding users' background or motivations to use an application could escalate a phenomenon known as "the rich-get-richer" [EK10] or even a cultural isolation of content, in which only popular items are shown on the top of recommendation rankings.

Due to the amount of content available online, information retrieval is the key method to the discovery of personalized information and recommendation of content whose goal is to identify users' preferences through their use and interaction with content and tools for resource evaluation. However, to grasp the relevance of each resource users interact with, an interface is necessary to obtain users' preferences by implicit [KT03, OLL08] or explicit [JSK10] feedback. Implicit feedback is

⁷Three of them were previously cited by us as an example of manual coding.

related to users' interaction with the system where they are not aware they are evaluating resources. For example, researchers infer that if users are clicking and visualizing content from a specific subject, maybe they would like to receive more resources from the same subject. Furthermore, tags, queries and comments made explicit are examples of implicit feedback, because users are not aware that these resources are used to collect their preferences and shape their profiles. On the other hand, explicit feedback is the evaluation users give to items through resources provided by the system interface as ratings, thumbs up/down, one to five stars, and so on. These are the approaches used for understanding users' preferences in recommender systems, and gathering user-data as source for recommending algorithms.

According to Adomavicius [AT05], recommender approaches can be divided mainly into collaborative filtering (CF), content-based (CB), and hybrid recommendation approaches. With the aim of explaining the ones we used in this work, we shall then detail CF and CB approaches, as follows.

- *Collaborative Filtering (CF)*: This recommender approach assumes that if two users rate n items similarly, the same users are more likely to share the same preference about other items [SK09]. The goal of this association is to find similar preferences among a large group of users to recommend interesting content. New item predictions are based on users' previous evaluations, gathered by the algorithm through search for users with similar evaluations. However, the downside regarding this approach is that most users do not perform item evaluations, making data for such analysis very sparse.
- *Content-based*: This approach [PB07] analyzes the content of a target user (one user at a time) to recommend new content to the same user. The advantage of this approach is that there is no need of other users' evaluation to recommend content. However, not having other users involved leads to the emergence of the phenomenon known as overspecialization, that is, the user is limited to receive recommendations that are always similar to those already rated because there is no information from other users' tastes to be compared with the target user's preferences and produce distinct recommendations.

2.2.1 Tagging Recommendation

Tag recommender systems have emerged to help users choose the most suitable tags to lead to better (more accurate, efficient, and satisfying) content retrieval. Recommendation of tags is especially useful for systems that count on huge amounts of content shared everyday. The manual classification of content is an impossible task to be conducted by experts. Due to that, user-generated tags or social tagging are conveniently used as a tool for indexing content on social media networks, for example. However, tagging can be a repetitive and tedious work, which demands attention, accuracy and counts exclusively on users' personal judgment to classify content. Therefore, recommender systems arose as a solution to improve tagging patterns, decrease users' effort to assign tags, and improve their quality.

Tagging recommendation approaches have been used and researched in many different contexts, such as enterprise applications [MKB⁺16], navigational tools to support users find content [SGMB08], framework for tagging recommendation [KKL17], tagging recommendation for photo sharing [dCZdO13, SvZ08], recommendation explanation [VSR09], tagging recommendation for microblogging posting [OWC14], and so on. In general, tagging recommendation approaches rely on tag quality, co-occurrence with other tags, popularity, object features, users' similarities, or the combination of document words and tags assigned in order to recommend other similar tags [DFT10, GLYH10].

In a previous work, the author of this thesis has developed a probabilistic model to recommend tags [dCZdO13, Z⁺12]. Thus, three measures were created based on co-occurrence, relevance, and popularity using a user-given tag (a query) to find other tags. This approach intended to improve tag quality by helping users find relevant tags based on topic similarity. The results found during a user experiment [dCZdO13] show that the homogeneity of tags increases as users start assigning recommended tags for the classification of the same image. This approach is used during the users' study we present on Section 3 in order to answer a hypothesis that led us to three research questions we aim to answer in the end of this work. Although recommendation is not the main focus of this work, this step was remarkable in the identification of patterns of tagging and how it is affected by this attempt of improving tagging homogeneity.

An important point to highlight is that regardless of the user's motivation, a model for recommending tags is developed by a designer that expresses his/her own point of view about similarities among users and/or tags for later recommendation. Figure 2.4 shows the three main parts that play a role in communication on recommender systems: the community of users from a social network (for example, Flickr, Instagram), which provides tagging vocabulary, representing a collective self-expression; designers, who communicate with users through an interface, and algorithms for recommending tags (using the data extracted from the community) according to the implemented approach; and the target user that uses tag recommendation and composes his/her tag vocabulary based on tags from other users. Once there is a system recommending tags, users can change the way they add them. They take appropriation from other users' tags at the time as they accept tag suggestions, beginning to model their own profile.

According to Nielsen [Nie12], many research projects have shown that developers, designers, and engineers assume that users' profile, behaviour, and needs are similar to their own. This type of assumption jeopardizes systems features, goals, and users' interaction needs, since it would result in a design that attends to designers/engineers/developers' own needs and not to the most important stakeholders of the system, the final users. Even when using collective knowledge as data source to generate recommendations, for instance, the analysis of users' behaviour should be addressed to draw users' expectations and motivations as a guide for system design. Next, we present the language theory used for the step of modeling tagging patterns (Section 5.1).

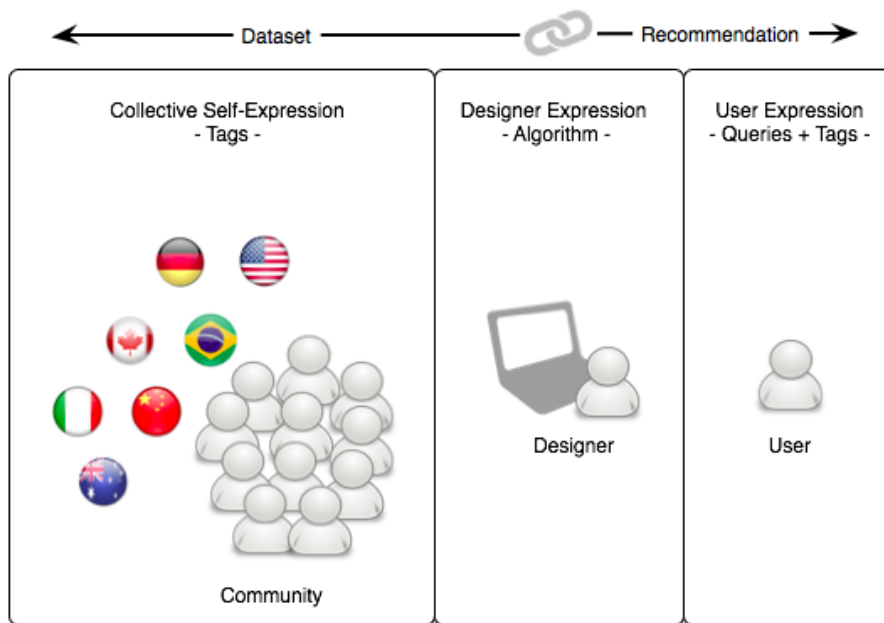


Figure 2.4: Differences among datasets, designers, and users.

2.3 Semiotic

Tagging systems features a multilanguage vocabulary: users may add tags from distinct languages to improve their photo retrieval on searches. The behaviour of tagging photos using more than one language to help search engines is considered good practice, but for some users tags in another language can be meaningless. According to Saussure, language is a social institution that comes from the individual history and it preexists based on a set of values [Net80]. Semiotic, also called semiology, aims to study the process of signification and communication of signs and symbols. It is considered part of social psychology, since the process of signification is personal, not static and it differs according to time and culture [Net80].

Saussure [DS11] described language as a system of signs that consists of a signifier and a signified. The signified aims to represent a concept, a content or the mental image it represents. On the other hand, the signifier represents the form the sign takes, for example, a word, an image, an illustration. Together, both aim to reach signification, which, in turn, is defined by the association between content and the expression used to represent it, and this association is related to culture [DS05].

Tagging systems have a process of attribution of signs (tags/words/sentences) by regular users according to their preferences, resources perception, language, culture, context, etc. In order to prepare a set of tags to be added to a resource, users must choose a tag (sign) among other ones from distinct languages, or new words that can be invented by a user or set of users. The meaning of signs arises from the difference between signifiers, which can be: syntagmatic (concerning position) or paradigmatic (concerning substitution) [Cha00, Net80]. The syntagmatic and paradigmatic relations provide a *structural* context where it is possible to categorize signs as codes:

- Paradigm is a set of associated signifiers, which differ from each other significantly. However, members from the same paradigm set can be replaced by others depending on the context they represent.
- Syntagm is a combination of signifiers that aim to form a meaningful and sequential text, whose elements may be related to each other. It is composed by two or more consecutive units.

Figure 2.5 shows the axes of the structure and the relation of paradigms and syntagms.

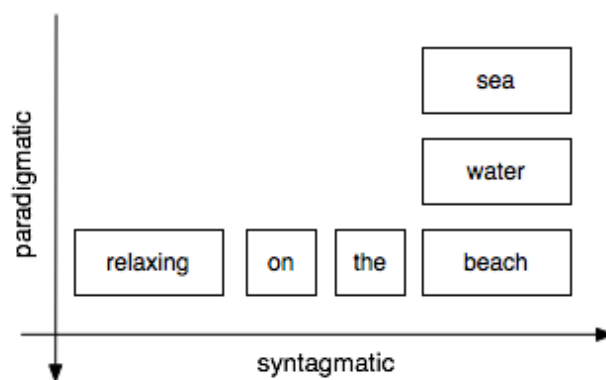


Figure 2.5: The dimensions of syntagmatic and paradigmatic relations.

Paradigms and syntagms can somehow be compared to a classical tag attribution behavior: distinct tags can have the same meaning and also sentences can represent the resource by a sequential combination of words that are associated to each other.

For Semiotic theory, the communication process uses the systems of signification and other codes or signs to achieve a variety of goals [DS05]. Tagging can be considered a way of communication and it also has a process of signification, once users have to interpret a text, a photo or a resource to add tags related to them or search using tags as queries.

Researchers [Rap02] have investigated the computation of word association to automatically retrieve words with syntagmatic and paradigmatic relation from a corpora. The use of co-occurrence, as cited before, refers to a word frequency computation not only used exclusively from recommender approaches, but also in the field of natural language processing (NLP) to extract pairs of words in a corpus⁸ with a combination of statistical generalization to predict which word combinations are more likely to appear in another corpus [Eve05]. They consider syntagmatic associations those words that frequently occur together, and paradigmatic associations words with high semantic similarity. Raupp was interested in the *association* of words in a Corpus. Differently, we are interested in the paradigmatic and syntagmatic dimensions as a *structure*, for example, the use of words as sequence (syntagmatic) to represent a tag and units (paradigmatic). More specifically, as stated by Saussure,

⁸Corpus is a set of writing text, being one of its application to extract the occurrence of words and combinations.

the syntagmatic structure is composed by elements presented in a sequence, forming a chain [DS11], a syntagm of words, a construction [Cha00]. This is the theory we used as basis for the investigation on how users assign tags and their motivation to do so.

3. USER STUDY

When users assign recommended tags, the set of their tags becomes more homogeneous. However, what has not yet been studied is whether the resulting homogeneity of tags also applies to the natural language adopted for tagging. At this point of our work it is important to highlight that although we have conducted this study using a recommender system as tool for understanding tagging behaviour, the results found here do not try to argue that one should or not use recommender systems to suggest tags. We decided to keep the use of a recommender system in one of the stages of the study to address modifications on how users behave and how this affects their tagging characteristics. We resort to mixed approaches for gathering data, and to conduct users' studies that gave us insight about users' behaviour on tagging systems. Results we found helped us compare how a recommender system could affect users' tagging from different points of view and gave us sufficient insight to create a model for tagging patterns. At this stage of the study we focused our analysis on quantitative data. To do so, we need to resort to different tools and statistics approaches to support our findings.

The study was conducted in two different countries, Brazil and Canada. Due to differences in the ethical committee rules regarding data gathering from one country to another, we conducted the qualitative data gathering using two different approaches, according to the requirements of each University. In both countries the studies we present here went through the ethical research committee for approval (Appendix A). Each participant was invited to read and sign a consent form that presents in details our research goals and the outcomes we expect to accomplish (Appendix B). In Brazil, during the stage of gathering qualitative data we approached users by conducting interviews in person, in which they answered open-ended questions while answers were being recorded. On the other hand, in Canada, the same questions were answered through a post experiment survey, where the questions regarding tagging patterns were inquired through two open-ended questions available in the same survey.

3.1 Method and Instrument Design

Since there were two main conditions to be investigated (tagging with and without the support of recommendation), participants were divided into two groups that had the experimental conditions changed: group one (G1) was asked to assign tags to photos with no recommendation support (NR), and then assign tags to the other set of photos in a different order supported by a recommender system (RS). Group G2 was exposed to the same conditions in the opposite order. As a general hypothesis, we first assumed that there is no difference in the use of tags from one stage to the other (with and without the support of recommendation). As results started to show differences among the use of tags from one state to the other, we proceeded to investigate more specific research questions, such as:

- Does the use of a recommender system change the tagging patterns (structure and language) users employ?
- Does tagging patterns change according to the class of photo being tagged, regardless of the system used?

The same methodology was applied for two different samples, to participants residing in a Portuguese speaking country – Brazil, and residents from an English speaking country – Canada.

In total, there were 57 participants from Brazil (26 female and 31 male, with a mean age of 26 years old), and 34 participants in Canada (19 female and 15 male, with a mean age of 25 years old).

3.1.1 Tagging using Recommendation

As a design platform, we used a model [dCZdO13] designed in a previous work by the author of this thesis, that has as its primary purpose the recommendation of tags. This model uses reference tags from users to recommend other tags (the so-called semi-automatic approach). In other words, after a participant assigns a tag he/she receives as recommendation a list of other tags that could be assigned to the same image based on co-occurrence. We instructed users on how it works and the options they could select after typing a reference tag:

“Every time you type a tag, the system will recommend other tags. You can select the tags you consider appropriate to the content being tagged”. Each user was asked to assign at least four tags to each image in the NR stage, and at least one tag in the RS stage.

We used a training dataset from Flickr for recommending tags with more than 600,000 tags. The dataset was acquired using the Flickr API. No prior choice of users/items/tags’ language was made while the tags were gathered. The utility of tags is computed by the combination of three measures for later combination to present a ranking of tags that can fit with a reference tag¹.

The recommendation model define each posting P_i as a triple $P_i = \langle u_i, r_i, T_i \rangle$ where $T_i = \{t_1, t_2 \dots t_n\}$ is a set of tags assigned to resource r_i posted by user u_i . This approach uses a reference tag t to get similar tags based on its co-occurrence in $P(t) = \{P_i | t \in T_i\}$.

Initially, it computes the k -tags with the largest co-occurrence from $P(t)$. A function records the existence of t in T and it is used to rank the co-occurring tags t_j by:

$$ranking(t, t_j) = \sum_{P_i \in P(t)} (t_j \in T_i) \quad (3.1)$$

After that, three measures are computed to penalize those tags that are not relevant or popular. These measures take from the top of the ranking tags used by few users but that are very frequent in the dataset. They are co-occurrence, relevance and popularity measures as following:

- *Co-occurrence* $coo(t, t_j)$: this measure is a normalization of the previous ranking. It computes both t and t_j by the number of items that have t , resulting in a value that can range from 0 to 1 for each t_j .

¹A tag typed by the users to the photo being tagged.

- *Relevance* $rel(t, t_j)$: this measure takes from the top of the ranking those tags that do not represent the community vocabulary, i.e., the name of the resource owner, tag reference to personal content, etc. It computes the number of users that used t and t_j by the number of items that have been found in the previous ranking by $ranking(t, t_j)$.
- *Popularity* $pop(t, t_j)$: this measure computes the popularity of t_j , that is, how important t_j is to the set of users that use t . The popularity is related to the frequency of use of t_j by the community. This measure uses the conditional probability as bases for computing the number of users using tags t and t_j divided by the number of users that use t .

Finally a ranking of recommended tags is computed by the geometric mean using the three previous measures.

3.1.2 Content-based classification of Images

To avoid biases on the classification of image category we used in the users' studies, we randomly recruited fourteen individuals (that did not take part in the main study reported in this work) to conduct the image content classification for the photos used in this study. Images used in this study were publicly available on Google Images and we report the links to them at Appendix C. As it is common in such research, images (Figure 3.1)² were classified by content [DF10].



Figure 3.1: Example of types of images used in this study.

Each participant received the images in a random sequence and were asked to classify the content of each image by its level of information presented regarding context: situation (whether the concept represented in the photo stands out) and location awareness (if the location where the photo was taken is obvious in the photo content). They also classified whether important parts of the image (the content) were clearly delineated from both the foreground and background. Table 3.1 shows the photo classification reported as having high level of content regarding the previously described classes. We named each one of the seven images as presented in the table.

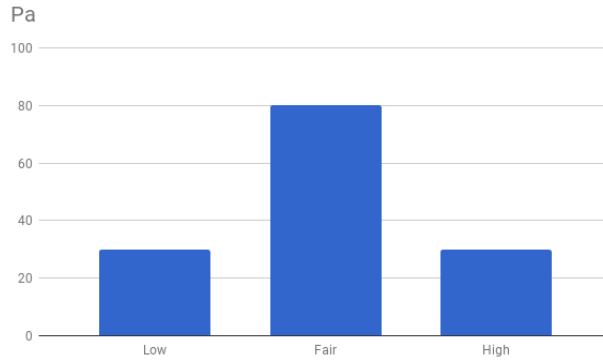
Regarding context relevance, we report the results according to the level (low, fair, high) of information classified by participants as context/situation prominence present in each image. For example, P_b and P_c were those images with most votes for context presence. The following Figures

²Images used in this study are publicly available for visualization online. However, we decided to only include links to them at the end of this work to avoid any concern regarding attribution of rights, since it was not possible to find information regarding the owner of some images.

Table 3.1: Results from content-oriented classification of photos.

Classification	Images						
	P_a	P_b	P_c	P_d	P_e	P_f	P_g
Prominence							
Background						×	
Foreground	×	×					×
Both			×	×	×		
Context (can be determined)							
Location					×	×	
Situation		×	×				

3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 report the results to each one of the images we use in the user studies we conduct in this work.

Figure 3.2: P_a classified as an image with fair context representation for its content.

3.1.3 Processing of Tagging Dimensions

In this work, we are interested in the use of words as units or sequences to represent a tag. To identify these differences, we modeled tagging from a semiotic point of view. This model was designed as we started conducting an open coding through the tags assigned by users. We stepped back to study about the possible structures used by users while assigning tags and their relation with semiotic structure. As we analyzed the tags, we found basically two main structures in the data that led us to conduct our data analysis following the language structure of syntagmatic and paradigmatic relations [DS11]. Any units or elements of language presented in sequence can be represented by a chain [DS11], as we have mentioned in Section 2.3.

The paradigmatic structure represents units assigned as tags. They are single words, putative tags used to describe objects, places, people. These are tags that most people would agree to assign, and are helpful to describe the content of an image. On the other hand, syntagmatic tags have a distinct structure, such as “*Just saying*”, “*Living my life*”, and users in general assign them to express more than a description of the resource explicit content. According to the definition of

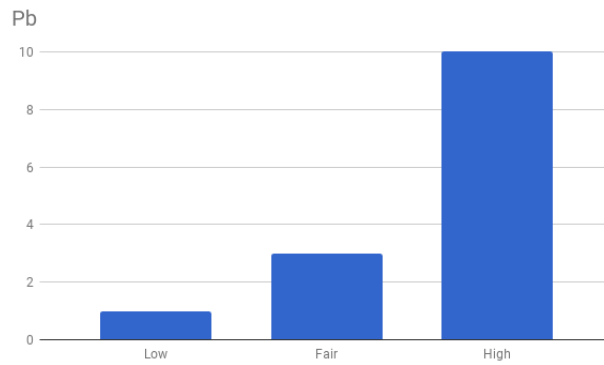


Figure 3.3: P_b classified as an image with high context representation for its content.

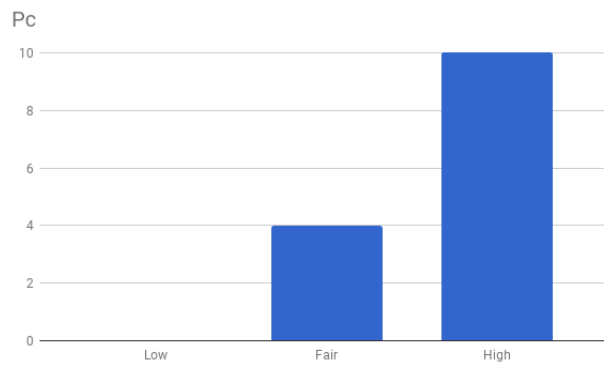


Figure 3.4: P_c classified as an image with high context representation for its content.

structure based on the Saussure chain, we use distinctive processes to quantify tag structures in each stage of this study.

The tags from the NR stage were manually coded as paradigmatic or syntagmatic once there was no difference among their source (all of them were added without the recommender system's assistance). In the RS stage, we coded tags using activity logs collected comprising information of all assigned tags, each one with an association about the original source (tags added by users as reference tags to get recommendation or recommended tags). This step in the process allows the observation and comparison of the frequency of reference tags against those tags that were recommended. We observed the long tail of power-law distribution of tags gathered in this study to classify the structure of each tag and, as we expected, the majority of syntagmatic tags were in the long tail. After the classification and frequency computation of each tag structure from both stages of the experiment, we focused our work on the statistical analysis of the data gathered.

3.1.4 Programmatically Classifying Languages

To process the language of tags assigned in this experiment, we used a standalone language identification tool based on a Naïve Bayes classifier [LB11]. This approach results in a probability estimation for a language when given a set of words. By performing the language identification

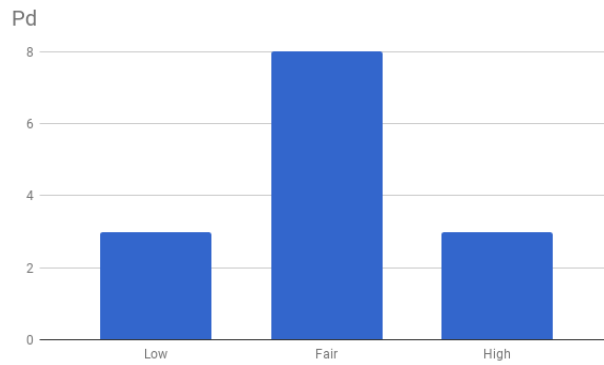


Figure 3.5: P_d classified as an image with fair context representation for its content.

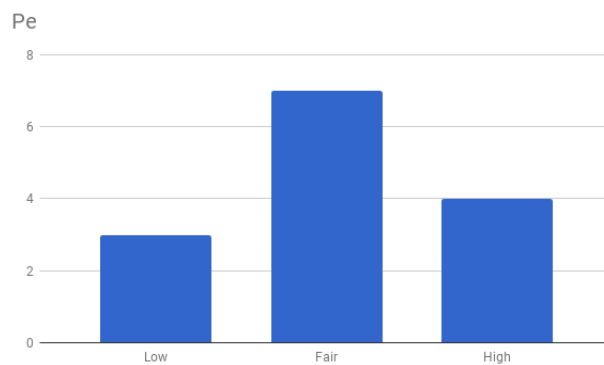


Figure 3.6: P_e classified as an image with fair context representation for its content regarding the image situation. It is important to point that this was an image that brings location as context information as well.

and observing a group of tags and its resulted probability estimation, we found that some users tagged photos multilingually, so the language classifier was useful to estimate a language score for each photo classifying it as mainly assigned with tags in English (EN) or Portuguese (PT), the two main languages used by participants in the tagging task. Also, we manually reviewed the language probability estimation to photos that presented proper names as tags e.g. *Eiffel* or tags without translation from one language to another e.g. *metro*.

3.2 Findings

3.2.1 Tagging Structure

In this work we do not argue that paradigmatic tags are better than syntagmatic ones. Instead, we aim to verify if users' tagging behaviour, regarding the tag structure, changes once they receive tag recommendation. To address our general research question for tagging patterns, we hypothesize that there is no relationship between the type of system used and the tag structure adopted. At first, we compared the tags assigned only to the same set of photos ($P_{a,b,c,d}$) presented in both stages of the experiment. For the study we have conducted in Brazil, participants assigned a total of 823

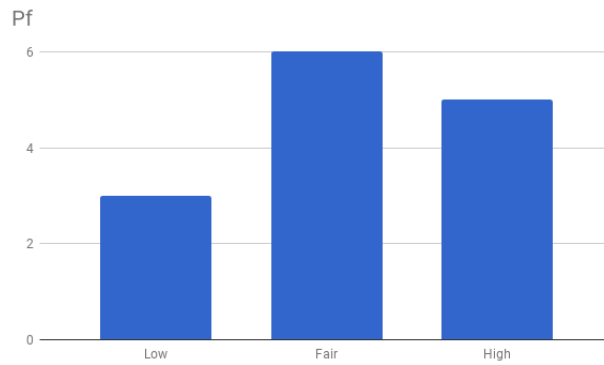


Figure 3.7: P_f classified as an image with fair context representation for its content regarding the image situation. It is important to point that this was an image that brings location as context information as well.

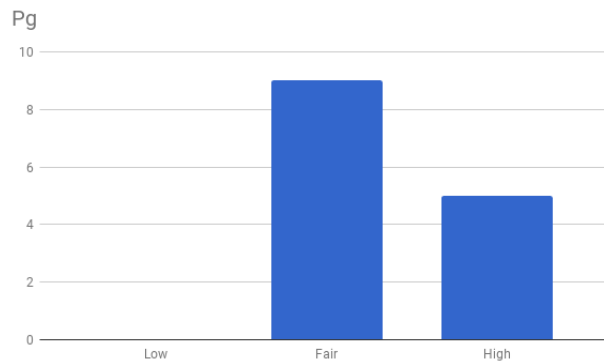


Figure 3.8: P_g classified as an image with fair context representation for its content.

tags in the RS stage for both G1 and G2. Table 3.2 summarizes the tag structures and their results from one stage to another. Results show that the proportion of syntagmatic tags changed when participants were aided by the recommender system. Results of a χ^2 test indicate an association among the variables for both groups.

Table 3.2: Classification of tag structures from participants from Brazil for the photos presented in both stage for each group. The p -value ($p < 0.01$) shows that there is an association between the type of system used and the type of tag structure assigned.

Tag Structure	G1		G2	
	NR	RS	NR	RS
Paradigmatic	479 (72%)	439 (94%)	355 (73%)	324 (91%)
Syntagmatic	193 (28%)	29 (6%)	132 (27%)	32 (9%)
Total	672	467	487	356

The same behaviour was observed for participants from Canada. When users were supported by recommendation their tagging patterns changed, according to the χ^2 result and the p -value reported in Table 3.3. They have assigned a total of 665 tags in the RS stage for both G1 and G2. At NR

stage we concluded that more tags with syntagmatic structure were assigned, while in the RS stage the proportion of paradigmatic tags had a significant increase.

Table 3.3: Classification of tag structures from participants from Canada for the photos presented in both stage for each group. The p -value ($p < 0.01$) shows that there is an association between the type of system used and the type of tag structure assigned.

Tag Structure	G1		G2	
	NR	RS	NR	RS
Paradigmatic	190 (59%)	285 (85%)	251 (67%)	289 (86%)
Syntagmatic	130 (41%)	43 (15%)	123 (33%)	48 (14%)
Total	320	328	374	337

One important aspect to analyze is that the photos chosen for this experiment highlight their content position, the context they represent (situation, concept or message standing out in the photo) and the context regarding the location where they were taken. Regarding image characteristics and differences, we conduct the next analysis with the null hypothesis in which the image class has no influence on the tagging patterns users choose for tagging. We compute the proportion of the use of each structure for each image according to the stage it was tagged. For this analysis, we expect that the proportion of syntagmatic and paradigmatic tags does not change regardless of the image class.

Table 3.4 shows the proportion of tags assigned in both stages (NR and RS) resulted from the tagging task from participants from Brazil. We found much evidence that photos with high associated context/situation are related to syntagmatic tags. In the NR stage, we found, mainly for G1, that the proportion of syntagmatic tags does not occur with the same proportion of paradigmatic tags to all photos, but P_b . Photo P_b stands out with its context and foreground objects. According to the p -value resulted from the z -test of proportion, its (P_b) proportion of syntagmatic and paradigmatic tags does not change significantly ($p > 0.05$). Differently, all other images presented a significant result regarding the proportion of syntagmatic tags ($p < 0.01$), even those photos ($P_{e,f,g}$) that were presented only in the RS stage still showed the same tagging behaviour of photos that were in both stages, which shows that the previous visualization of photos did not influence the tagging task in this study.

On the other hand, when we look for the results we found for participants from Canada, we noticed that they are more inclined to assign tags with syntagmatic structure. We found evidence for G1 group that P_a, P_b images users are more likely to assign syntagmatic tags during the NR stage (z -test p -value > 0.05). However, the same behaviour was not observed for G2. This could indicate that the RS stage and the recommendation results may influence the way users assign tags during the NR stage. Mainly in those cases where it was possible to find more tags that describe the image content instead of making reference to the context it represents. To verify the tagging difference from one stage to another, we compared the tag structure proportion from one stage to another through the z -test of proportion. When participants were aided by recommendation, their

Table 3.4: Proportion of syntagmatic and paradigmatic tags in both stages of the experiment to each image assigned by participants from Brazil.

	Stage	Tag	P_a	P_b	P_c	P_d	P_e	P_f	P_g
G1	NR	Syntag.	0.29	0.45	0.20	0.23	-	-	-
		Parad.	0.71	0.55	0.80	0.77	-	-	-
	RS	Syntag.	0.04	0.12	0.02	0.05	0.04	0.05	0.04
		Parad.	0.96	0.88	0.98	0.95	0.96	0.95	0.96
G2	NR	Syntag.	0.32	0.39	0.16	0.23	-	-	-
		Parad.	0.68	0.61	0.84	0.77	-	-	-
	RS	Syntag.	0.07	0.18	0.07	0.05	0.02	0.03	0.06
		Parad.	0.93	0.82	0.93	0.95	0.98	0.97	0.94

Table 3.5: Proportion of syntagmatic and paradigmatic tags in both stages of the experiment to each photo (Canada).

	Stage	Tag	P_a	P_b	P_c	P_d	P_e	P_f	P_g
G1	NR	Syntag.	0.50	0.58	0.24	0.31	-	-	-
		Parad.	0.50	0.43	0.76	0.69	-	-	-
	RS	Syntag.	0.13	0.27	0.04	0.12	0.10	0.10	0.08
		Parad.	0.87	0.73	0.96	0.88	0.90	0.90	0.92
G2	NR	Syntag.	0.31	0.44	0.31	0.26	-	-	-
		Parad.	0.69	0.56	0.69	0.74	-	-	-
	RS	Syntag.	0.08	0.24	0.13	0.10	0.08	0.05	0.06
		Parad.	0.92	0.76	0.87	0.90	0.92	0.95	0.94

tagging behaviour changed ($p < 0.01$) from one stage to another. This effect occurred also to P_b , which presented, mainly for G1, no difference in proportion of syntagmatic and paradigmatic tags.

Moreover, to illustrate the differences between the vocabulary agreement on both stages and the type of tag assigned, Figure 3.9 and Figure 3.10 (A) shows the distribution of tags in the NR vs. RS stage, and (B) the distribution of reference and recommended tags in the RS stage from participants from Brazil and Canada, respectively. We extracted the tags that were in the head and in the long-tail of the power-law distribution of tags. First, we looked at more frequently assigned tags in the NR stage (A). The head of the power-law, that represents the common vocabulary of participants, is represented by 58% of tags assigned by participants from Brazil, and a similar proportion was found in the data assigned by participants from Canada, which had 54% of tags assigned more than once. Most frequently assigned tags had paradigmatic structure, and were used to describe photo content as *camping, vacation, beach* [Canada]; *férias³, praia, camping* [Brazil]. Surprisingly, they are the exact translation from one another.

In the head of the power-law, 6% of tags had syntagmatic structure, for both datasets. On the other hand, in the long tail, represented by (Brazil 42%; Canada 46%) tags that were assigned only once, 57% of them had syntagmatic structure in the results from Brazil, and 62% represented in the results from Canada. The syntagmatic tags found seem to be motivated by social communication,

³férias = vacation and praia = beach.

self-expression (opinions, emotions) and personal tags: *happymonday*, *funnyday*, *crazyexperience*, *nosensefriend[sic]*, *mypet*.

On the other hand, when comparing the power-law distribution from the RS stage, tags assigned had a higher agreement than the NR stage. The head of the power-law of the RS stage represents 77% of tags and the long tail 23%.

Looking at the power-law distribution of the type of tags of RS stage (Figure 3.9 (B)), we found that 62% of tags were assigned as reference tags and 48% were assigned based on recommendation (Brazil). Compared to Canada, there was no significant difference among their tagging behaviour during the RS stage: 61% assigned reference tags as 49% based on recommendation (Figure 3.10 (B)).

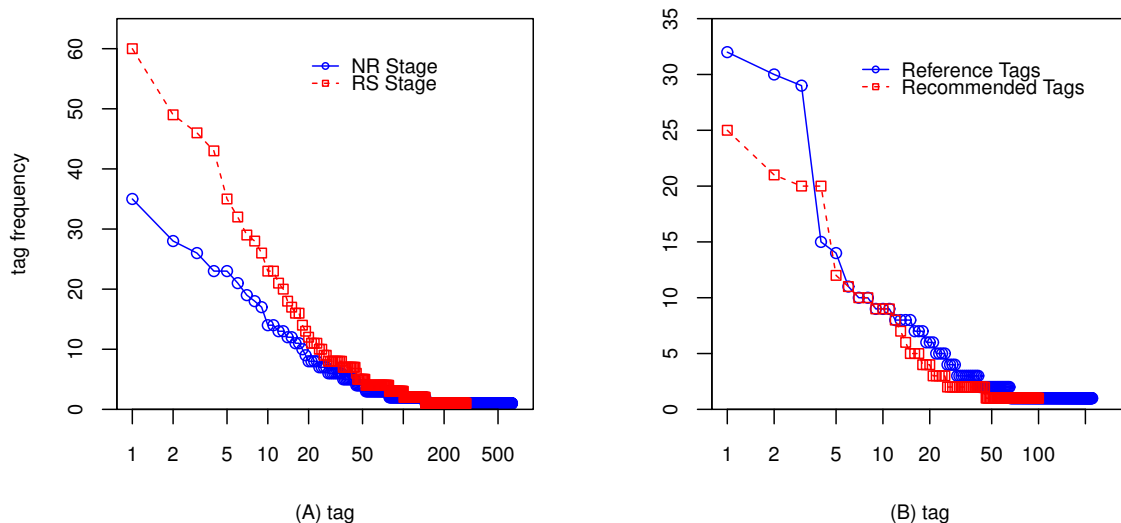


Figure 3.9: Power-law distribution comparing NR vs. RS stage (A) and the distribution of reference and recommended tags (B) from Brazil.

From the set of reference tags, 90% were paradigmatic and 10% syntagmatic in which 87% were in the long tail: *supercute*, *dontfallasleepatthesubway* (Brazil). Compared to Canada results, it seems that participants are more inclined to use syntagmatics as reference tags than participants from Brazil. Results show that in the RS stage 19% of reference tags had syntagmatic structure against 81% of paradigmatic structure. From this set, 82% were in the long tail.

When we look at the list of recommended tags assigned, syntagmatic tags represented only 2% of recommended tags, and 65% of them were in the long tail from data gathered from Brazil: *morningafter*, *familyvacation*. Also, a similar behaviour was found from data collected in Canada: 1% of syntagmatic structure and 99% paradigmatic. From the set of syntagmatic tags, 72% of them were in the long tail – *workcolleagues*, *wildlife*.

Syntagmatic tags in the head of the power-law in the RS stage seem to be more related to content description (*sunnyday*, *blueeyes*), while in the NR stage, besides the syntagmatic tags in the

head used to content description *whitecat*, *blueseas*, tags were also found for social communication and related to photo context (*loveit*, *bestfriends*, *bestpicture*). This tagging behavior suggests that there may be a difference in syntagmatic tags that are from common agreement from those that are not. This observation opens space for future investigation since the quantity of syntagmatic tags resulted from our study is not enough to generalize a conclusion for the vocabulary commonly used for this type of tag.

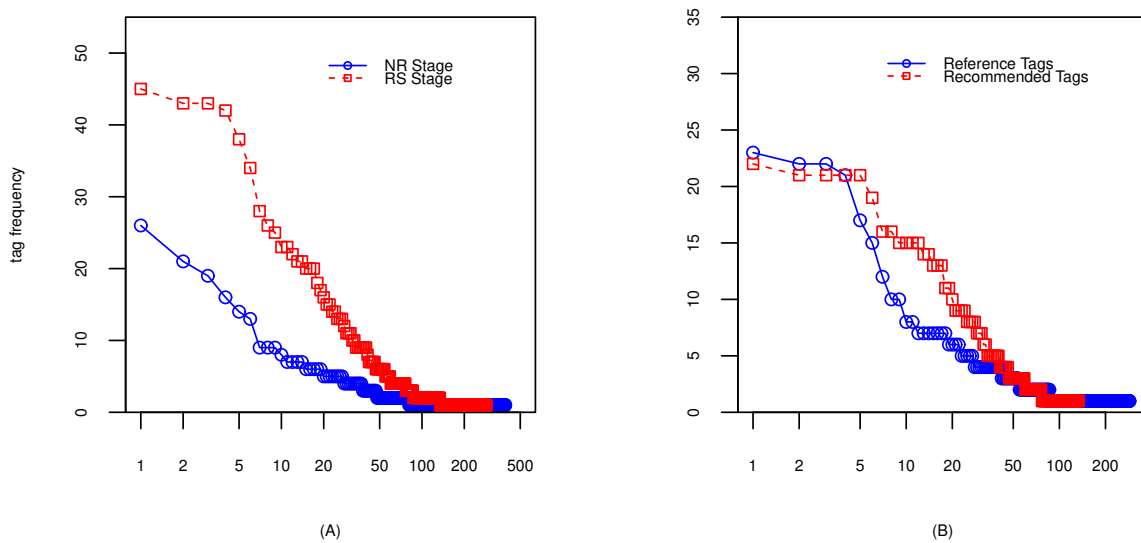


Figure 3.10: Power-law distribution comparing NR vs. RS stage (A) and the distribution of reference and recommended tags (B) from Canada.

3.2.2 Language

In this stage, we focused on the analysis of the most frequently adopted language in each stage of the experiment. We divided the results into two sections to discuss the differences we found according to Portuguese and English speakers. To address our general research questions for tagging patterns, at this stage of the analysis we assume that there are not differences in the language adopted from NR to RS stage.

English speakers

In this stage of the experiment, we found that during the NR stage, all images for both groups G1 and G2 were mainly assigned with tags in English. The choice of language for indexing content online is a practice that puts resources in a pool of content regarding the subject matter they represent. In the work of Ronen et al. [RGH⁺14], they found that English is a language that works as a hub to connect content and has halo composed by other languages such as German, French,

and Spanish. What we found in the results of tags assigned by participants from Canada is that they only assigned tags in English during the NR stage, even the country having two official languages, French and English. During this stage we conducted a manual analysis, since the volume of tags were smaller compared to the data we collected from Portuguese speaker participants. The only difference we found was for tags assigned for the P_f image during RS stage. As we mentioned in the image classification section, this is an image that has a clear cultural icon indicating the location where it belongs. This image was used only during the recommendation stage and was the only one that presented tags assigned in a language other than English. The tags were recommended in French based on a reference tag assigned in English – *eiffeltower*. In total 13 tags were assigned in French following the recommendation suggestion – “*toureiffel*” – and other 2 were assigned as reference tag. All tags were assigned by different users, meaning that 15 users in total presented this behaviour. Because this tagging pattern only occurred in the RS stage and is associate to an image that has reference to its location, we believe tags that co-occur with location-based tags are more likely to occur in English.

In order to verify how participants assign tags and the classes that were more frequently assigned in the head of the power-law, we present in Table 3.6 the most popular tags used in both stages of the study. We used the classes of tags reported by Gupta et al. [GLYH10] to identify the tags assigned as: Content-based (CB) — tags describe the content in the photo (e.g., *towel*, *plaid*, *grass*); Proper-names (PN) e.g., *Great Wall*; Subjective Tags (ST) which express users opinion or emotion (e.g., *beautiful* and *sad*). Also of note is that participants assigned acronyms-based tags (AT), e.g, *bff*, *yolo*⁴.

Although these classes cover a wide variation of tags, we included a class called *Concept-based tags* (CP) used for classifying the image concept. Concept-based⁵ tags express the mental combination of the photo characteristics which result in tags such as *busy*, *cold* and *helping*. The way users assign tags to photos is different from the way they tag text. An image brings more than a collection of content/objects, it can bring a combination of elements that can be expressed e.g. by a concept. We used the term concept-based and location-based (LB) terms instead of context-based (previously reported as a class of tag related to location/time [GLYH10]) to avoid ambiguity of terms for classes of tags that are so distinct from each other. Contextual tags were not presented in the top of the ranking of tags more frequently assigned but we considered this class as tags, such as syntagmatic tags – “*Best day ever*”, “*Friends forever*” – that represent more than the image content, and bring sometimes implicit information that are related with the context or situation during the time the picture was taken.

Regarding the type of tags assigned, as previously stated, the RS stage differs by how users assigned tags, that is, as reference tags or tags that participants accepted by recommendation.

⁴‘bff’ = best friends forever, ‘yolo’ = you only live once.

⁵Note that the CP tags are different from tags in the ST class which cover tags that represent opinion and emotion.

Table 3.6 shows the differences and the frequency of tags assigned among both groups in different stages.

This behaviour will be better addressed in the analysis we conducted for the open-ended questions participants answered after the NR and RS stage, which we discuss on Section 4.

Table 3.6: The list of most frequently assigned tags, their class and the representative proportion among the set of tags of the same photo*.

Stage	Photos										
	P_a	P_b	P_c	P_d	P_e	P_f	P_g				
NR	cute	ST	subway	CB	camping	CP	beach	CB			
	kitten	CB	sleeping	CP	friends	CB	vacation	CP			
	kitty	CB	peace	CP	outdoors	CB	relax	CP	-	-	
	ball	CB	funny	ST	nature	CB	sun	CB			
	cat	CB	passedout	CP	campfire	CB	relaxing	CP			
		28%		17%		38%		34%			
RS	kitten	CB	metro	CB	nature	CB	beach	CB	statueofliberty	PN	
	kitty	CB	tired	CP	camping	CP	ocean	CB	nyc	LB	
	cute	ST	subway	CB	hiking	CP	vacation	CP	newyork	LB	
	cat	CB	sleepy	CP	campfire	CB	sea	CB	newyorkcity	LB	
	pet	CB	transit	CB	friends	CB	travel	CP	manhattan	LB	
		45%		24%		48%		44%		49%	
									paris	LB	
									eiffeltower	PN	
									france	LB	
									europa	LB	
									snow	CB	
										pet	CB
										46%	

According to the results we found, similar class of tags were found in both stages of the experiment, but the proportion that it represents from one stage to another shows how the homogeneity of tags change as users are guided by a recommender system. During this stage, P_b , one of the images with the higher value for context and considered with foreground prominence, presented more variability among the choice of classes for tagging. It presented three class of tags – CB, CP and ST – and this behaviour changed for the tags more frequently assigned in the RS stage. Participants assigned more tags to describe the content presented in the image, using content-based tags. Also, the agreement among the use of CB tags were higher in the second stage when comparing the results from P_a . It increased 17% in the ranking of tags more frequently assigned from one stage to another. This are paradigmatic tags, as we have mentioned before, these are tags considered putative, and of high agreement among users.

The images that were presented in the second stage, two of them that had cultural icons involved were those that had location-based tags as participants favorite. Although they had more than one unit to represent the content, we did not considered them as syntagmatic tags because the context association is related to location and not image situation.

Portuguese Speakers

Differently from English speakers participants, Portuguese speakers presented a different behaviour regarding the use of tags in foreign languages [ZBS16b]. In this stage, because of the volume of tags that were assigned in other language different than Portuguese we classified lan-

guage using the standalone language identification tool that results in a probability estimation based on a preset language to each one of the images assigned by each user.

The set of tags assigned to each image by each participant was classified as either PT or EN. Table 3.7 shows the difference in the proportion of images and the main language used by each group of images that were presented in both stages of this study.

Table 3.7: Comparing the proportion of images assigned mainly in PT and EN in both stage of the experiments.

	G1		G2	
	NR	RS	NR	RS
PT	81 (61%)	24 (18%)	63 (65%)	25 (26%)
EN	51 (39%)	108 (82%)	33 (35%)	71 (74%)

When not using the recommender system, participants tagged fewer images using EN (G1: mean = 1.54 SD = 1.60; G2: mean = 1.37, SD = 1.71). However in the RS stage, more images were tagged mainly in EN (G1: mean = 3.27, SD = 1.30; G2: mean = 2.91, SD = 1.28). A (paired) Wilcoxon signed-rank test indicated that the mean of images with tags in EN changed ($p < 0.01$) from one stage to another for both groups. This behaviour also was found when we looked to the language used in each image individually, before and after recommendation (McNemar $p < 0.01$) and also for the images that were tagged only in the RS stage.

To make sure that the results found in this study were not narrowed by a few participants' behaviour, we looked at their results individually. We classified users as: **PT-taggers, EN-taggers or multilingual-taggers (ML-taggers)**; PT-taggers — had all their images classified mainly by tags assigned in PT; and ML-taggers had a mix of images tagged in EN and PT. Figure 3.11 shows the proportion of participants and their respective tagging behaviour in each stage of this study during the experiment conducted with Portuguese speakers.

At the NR stage, 45% of participants were classified as PT taggers. However, this behaviour changed in the RS stage, only 8% of them kept tagging images mainly with tags in PT. In the RS stage, the majority of PT taggers switched their tagging language and behaved as EN- and ML-taggers.

To try understand participants' behaviour, we examined the order of tags assigned in the RS stage: We noticed that at first some images received reference tags in PT but the following reference tags were assigned in EN. We hypothesized that, as participants received tag recommendation in EN, they switched the language of reference tags. However, individual users' behaviour needs future investigation.

In order to investigate the class of photos and its relation with language when tagging is supported by recommendation we looked to the content-classification of photos. Figure 3.12 shows the proportion of photos with tags mainly in EN or PT for the photos that were presented in the two stages of the study. The z -test of proportions shows that the proportion of PT and EN to each photo did not change ($p > 0.05$ for all) when the recommender was not used (NR).

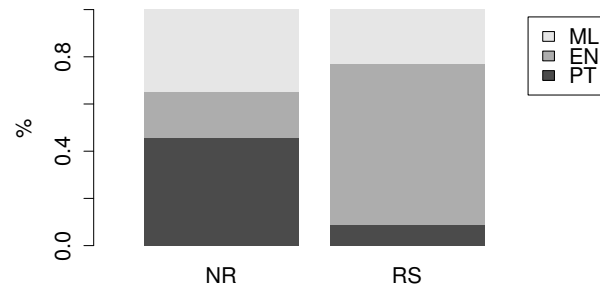


Figure 3.11: Comparing participants' language chosen for tagging in both NR and RS stages.

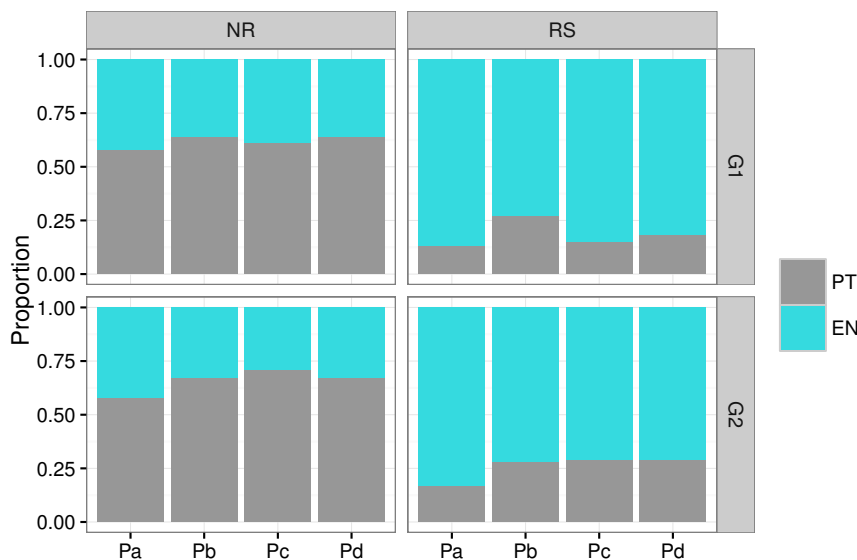


Figure 3.12: Languages adopted on photos presented in both stages.

However, a distinct behaviour was observed in the RS stage, the proportion of photos with tags assigned mainly in EN did increase in both groups. A McNemar paired test showed that the language used for tagging differed ($p < 0.01$) when using recommendation and not.

Moreover, regarding the remaining photos presented only in the RS stage, we found particular evidence that the location⁶ of the photo affected the language used for tagging. Among the high proportion of tags in EN assigned to photo P_f , it also had tags in French. The tag “*tou Eiffel*” was recommended by the system and it was the only tag in French that was assigned more than once: 35% of participants accepted this recommendation. Three participants also assigned reference tags in French (*vive la France*, *bon voyage*[sic], *bonvine*, *froid*) for P_f but it was not enough to classify French as a language that stood out among EN and PT.

⁶Based on the previous content-classification of P_e and P_f

P_g , that was also presented only in the RS stage, showed a similar language behaviour compared to P_a in the RS stage: They were both classified mainly by their foreground content and the proportion of tags in EN was significantly higher than PT (z -test $p < 0.01$). To understand the relationship among the language and the type of tag assigned to each class of photo, we turned to the list of tags gathered in this study (see Table 3.8). It shows those tags that were more frequently assigned, their class and their proportion of these sets represent among other tags assigned to the same photo.

Table 3.8: The list of most frequently assigned tags, their class and the representative proportion among the set of tags of the same photo*.

Cond.	Lang.	Photos							
		P_a	P_b	P_c	P_d	P_e	P_f	P_g	
NR	PT	gato <i>CB</i>	metro <i>CB</i>	acampamento <i>CB</i>	praia <i>CB</i>				
		fofo <i>ST</i>	sono <i>CP</i>	amigos <i>CB</i>	férias <i>CP</i>				
		gatinho <i>CB</i>	zoeira <i>CP</i>	fogueira <i>CB</i>	sol <i>CB</i>				
		brincando <i>CP</i>	dormindo <i>CP</i>	férias <i>CP</i>	mar <i>CB</i>	-	-	-	
		filhote <i>CB</i>	mico <i>ST</i>	natureza <i>CB</i>	verão <i>CP</i>				
		15%	15%	24%	27%				
NR	EN	cat <i>CB</i>	sleeping <i>CP</i>	camping <i>CP</i>	beach <i>CB</i>				
		cute <i>ST</i>	selfie <i>CP</i>	friends <i>CB</i>	vacation <i>CP</i>				
		kitten <i>CB</i>	subway <i>CB</i>	fire <i>CB</i>	summer <i>CP</i>				
		kitty <i>CB</i>	funny <i>ST</i>	trip <i>CP</i>	sea <i>CB</i>	-	-	-	
		pet <i>CB</i>	lol (AT)	cold <i>CP</i>	relax <i>CP</i>				
		18%	9%	18%	16%				
RS	PT	animal <i>CB</i>	metro <i>CB</i>	acampamento <i>CB</i>	praia <i>CB</i>	liberdade <i>CP</i>	paris <i>LB</i>	animal <i>CB</i>	
		gato <i>CB</i>	trem <i>CB</i>	fogueira <i>CB</i>	férias <i>CP</i>	estatuadaliberdade <i>PN</i>	metro <i>CB</i>	dormindo <i>CP</i>	
		gata <i>CB</i>	dormindo <i>CP</i>	natureza <i>CB</i>	sol <i>CB</i>	america <i>LB</i>	metro <i>CB</i>	cachorro <i>CB</i>	
		filhote <i>CB</i>	sono <i>CP</i>	amigos <i>CB</i>	verão <i>CP</i>	férias <i>CP</i>	fran ca <i>LB</i>	amor <i>ST</i>	
		fofo <i>ST</i>		acampando <i>CP</i>	mar <i>CB</i>	turismo <i>CP</i>	frio <i>CP</i>	labrador <i>CB</i>	
		13%	12%	9%	12%	10%	21%	7%	
RS	EN	cat <i>CB</i>	subway <i>CB</i>	camping <i>CP</i>	beach <i>CB</i>	statueofliberty <i>PN</i>	eiffeltower <i>PN</i>	dog <i>CB</i>	
		kitty <i>CB</i>	sleep <i>CP</i>	nature <i>CB</i>	sea <i>CB</i>	newyork <i>LB</i>	france <i>LB</i>	cute <i>ST</i>	
		kitten <i>CB</i>	sleepy <i>CP</i>	friends <i>CB</i>	ocean <i>CB</i>	usa <i>LB</i>	europe <i>LB</i>	baby <i>CB</i>	
		cute <i>ST</i>	tired <i>CP</i>	vacation <i>CP</i>	vacation <i>CP</i>	nyc <i>LB</i>	tower <i>CB</i>	puppy <i>CB</i>	
		pet <i>CB</i>	sleeping <i>CP</i>	hiking <i>CP</i>	summer <i>CP</i>	ny <i>LB</i>	snow <i>CB</i>	pet <i>CB</i>	
		52%	23%	42%	41%	48%	26%	54%	

* Where the list has fewer than five tags, participants did not assign other tags in that language more than once.

The more noteworthy result was observed for P_b that was previously content-classified as high context regarding the photo situation. This photo shows a wide type of tags variation in EN in the stage without the recommendation support (NR). This behaviour completely changed in the RS stage, the CP class in EN stood out among other classes. We looked to the type of tag assigned to this photo in the RS stage and we found that from this list that represent 23% of tags assigned, only 25% were tags recommended by the system. This result shows that the tags assigned to this photo do not represent a high general consensus among participants. We compare this tags to the tags assigned to P_c , a photo that was also classified by its context (situation) and we found that the tags used differ from P_b mainly because P_c had more CB tags reflecting the content-classification made to P_c (back/foreground in the same level of prominence). Also, P_c and P_d had both the same content-classification regarding prominence, and similar tagging behaviour in the proportion and class of tags assigned using PT in the NR stage. The set of tags in PT more frequently assigned to P_d was the same in both stages (NR and RS). The proportion of this set in PT has dropped in

the RS stage to less than half and only 4% of these tags were assigned using recommendation. A similar behaviour was also observed on tags in EN assigned to P_a . Although the same set of tags in EN is present in both stages, in the RS stage this set represents more than half of tags gathered to P_a , showing that the tag homogeneity occurs alongside the language adopted.

Another interesting observation was the type of tag assigned to P_e and P_f . The tags in EN assigned to P_e represent almost half (47%) of the tags assigned to this photo. Most class of tags in EN were location-based, whereas in PT only one tag was from this same class. Although this behaviour was expected because of the class of the photo (high context related to location) it stood out mainly in EN. We also turned to the type of tags assigned and different from the behaviour we found on P_b , 74% of this set of tags in EN were recommended by the system.

Although P_f also had the same content-classification from P_e and the proportion of tags in EN assigned to P_f was significantly higher than PT in the analysis we previously showed, most frequent tags assigned had similar proportion in EN and PT. The main difference between the tags assigned to P_e and P_f is reflected by their content. P_f was content-classified by its prominence of the background and it received much attention as the location of the photo, reflecting it in the class of tags assigned. P_f had more CB than LB tags compared to P_e . It is important to notice that tags like *Paris*, *Eiffel* and *metro*, that are in the list of tags in PT, are words that have no difference of translation from PT to EN and they had impact in the proportion of tags in PT. Also, among the tags more frequently assigned, there was a tag in French (*tour Eiffel*), that represent 3% of tags assigned to P_f .

3.3 Discussion

Although we have identified how a recommender system affects tagging behaviour, the main goal of this work is to understand which are the differences regarding tagging patterns and motivations for doing so.

Among the differences we found, the structure and the language were the main indicators that pointed that users will choose different tagging patterns according to their goals or type of image they are tagging. Images that have high context involved are more likely to receive tags that have syntagmatic structure. These types of tags showed lower agreement among participants, and were mainly present in the long tail of both stages of the experiment and for all groups of participants. Syntagmatic tags indicate they might have an association with tags that represent the context of the image. Participants will assign it as a tool for sending a message that is not related to the image content per se. For example, when assigning tags to the P_b image, participants used a wide variation of tags that are more related with self-expression than the objects and elements presented in the image. Tags such as, *#notimpressed*, *#studentlife*, *#zoeiraneverends*⁷, show that each participant has a different interpretation of the image situation, and so, it will present a lot of variability in the set of tags assigned to the image. However, this behavior was not so evident when participants

⁷A tag that mixes Portuguese and English.

assigned tags supported by recommendation. Participants assigned more paradigmatic tags in the second stage of the experiment compared to tags from the first stage. Regardless of the class of the photo being tagged, participants changed the type of tag structure adopted when supported by recommendation.

In contrast to syntagmatic tags, the paradigmatic tags had a higher agreement among participants in both stages. It stands out in the RS stage, and it affects also the language used for tagging among participants from Brazil. Those who already had a tendency to assign tags in English, had this tagging pattern highlighted during the second stage. The proportion of images assigned using English from one stage to another shows the influence of tagging popularity in the pattern of tags users choose.

We found that recommender systems could affect users tagging patterns from a structural perspective and language choices. This result strengthens the need for a model to support the identification of tagging patterns and motivations, and a framework to support the identification of tagging behaviour by users profile.

4. UNDERSTANDING USERS' TAGGING PATTERNS

At this stage of the research, we focus on gathering and analyzing qualitative data to better address why users choose to assign tags using different languages and structures [ZMS]. We resorted to open-ended questions that could support data gathered in the users' experiment.

4.1 Method and Instrument Design

After analyzing the quantitative data from the experiment conducted with participants from Brazil, we decided to resort to qualitative research to better address the reasons that lead participants to choose among distinct structures and languages for tagging. We aim at distinguishing tagging preferences among users, and addressing what are the reasons for them to switch from one tagging pattern to another.

About six months after the experiment, participants from Brazil were contacted to take part in the second stage [ZMS] of our research to address answers related to the results we have found in the quantitative stage of the studies. At this stage, a total of six participants (three females and three males, with age ranging from 21 to 35), took part in this stage to answer a set of questions during an interview. We conducted a semi-structured interview using open-ended questions that have emerged from the analysis and observations of the data collected during the first study. The interview was structured to investigate aspects about tagging choice criteria. During the interview, we asked questions about the participants' choice for tagging regarding language and structure:

- Do you use foreign languages to assign tags? Why?
- Do you use tags with more than one word? Why?

Due to limitation of resources and ethics requirements to collect and use data for conducting research abroad¹, participants from Canada were asked to answer these same questions related to the same subject, but during the demographic survey application. 13 of them answered the same two open-ended questions regarding the choice of the language and structure for tagging.

In order to analyze the data obtained in this study, we applied Grounded Theory [CS08], that postulates open coding and axial coding as main tools to build up concepts about the gathered data. Three researchers led this work, in which each one of them conducted their own open and axial coding in order to define concepts about the data. In the open coding step, the focus was on the analysis of the transcribed interviews to identify some categories that could help us answer our research questions. Thereon, during the axial coding step, it was possible to identify relationships among the categories generated by the previous step (open coding). Based on this analysis, each researcher

¹There are specific requirements in Canada referring to users' data confidentiality and how they can be used on research that intends to take data collected from Canadian residents to another country.

generated his/her own concepts about the participants' answers. Thereafter, a triangulation was conducted among the concepts found by each researcher, resulting in main concepts and sub-concepts that will be discussed in the following section alongside some participants' sentences said during the semi-structured interview.

4.2 Results

4.2.1 Language Choice

Regarding the question about the reasons for tagging in a foreign language, results show three main concepts: *routine* (for both groups of participants from Canada and Brazil), *indexing* (Brazil) and *contextualization* (Canada). These were the concepts found during the quantitative analysis using Grounded Theory tools to build the conclusions according to participants' answers.

Portuguese speakers reported that the decision on language choice for tagging is associated with their everyday life. The contact with foreign content is what builds their tagging patterns. P2: *"The English language, for example, is part of my daily life. I read a lot of articles in English... it is not a matter of preference. I think it is now part of my personality."*

For this participant, the use of tags in another language is guided by his routine, and the daily contact with resources from foreign languages seems to be the main influence on his tagging language choices. This was reported by P1, as well. However, differently from P2, his main choice for tagging was Portuguese language, and he expressed that he feels safer to assign tags using his mother language: P1: *(...) I add tags in Portuguese, this is my personal insight. I do not feel creative to assign tags (...) So, when I assign them, I use Portuguese.*

Moreover, participants mentioned they are aware of the possibilities to spread content using tags in other languages. The propagation of content was one of the motivations they mentioned. This pattern of tagging is used as a tool to spread content, which means that content can be indexed and reach expanded audiences. These are patterns of tags that can be used to communicate with more people, even with unknown users, or those who are not part of their personal network (followers, friends, etc.). It shows evidence that language choice is associated with the audience they intend to reach.

We could observe that to P4 the task of assigning tags seems to be a safe way for self-expression and spreading of content, even using another language. P4: *"(...) for example, on Instagram, tags reach more people. Besides that, I do not like to write a lot. It is easier to express myself when I use tags."* Although he is not an English native speaker, he can express himself by tags and increase his content visibility among people that use key words in English for searching online content. This could be related to the fact that instead of writing a lot of words in a foreign language, users can express themselves in a foreign language by using few words (even syntagmatic tags) and, yet, expand their audience. Tags in English are seen as a safe way to self-expression, and are still able to reach a broad audience using pools of tags, or memes, for example, that are very popular and promote context about an image and its topic.

Following, we have P5's point of view regarding the use of foreign language for tagging. She stated that she knows exactly what tags are for. She knows that, when using tagging, the indexing goals are associated to it and the choice of using tags in a foreign language is related to the fact that more people will have access to the resource being tagged. P5: "(...) *in English, for example, you can reach a higher number of people, because English is one of the most spoken languages.*"

Choosing another language is a tool used for propagation, an *indexing* strategy to reach a broad *audience*, increasing the number of likes an image can get. They also try to reach distinct audiences through some popular/funny expression in a foreign language (e.g. "memes"²).

The use of "memes" was also mentioned by **English Speakers** as a way to spread an image to different audiences. However, differently from Portuguese speakers, they did not mention the use of foreign languages when tagging with indexing purposes. What we have found is that English speaker participants are not concerned about whether their message is been spread to another audience according to the language used. Our insight about this behaviour, based on their answers, was that they have this feeling that their target audience is "listening" to them because they assign tags in English. As P6 says: "*I don't use a foreign language to assign tags because the majority of people who live in Canada speak English.*" For him, the language is not seen as a tool for spreading content, he is more concerned about his audience. Many of them also mentioned concerns about their language mastery. We considered this behaviour as part of the routine concept as well. P1: "*I mainly use English because I am most comfortable with it. However, I do use French when describing something very nice or fancy.*" Participants mentioned that if they do not know enough of the language to express themselves, they will avoid to use it in the tagging task. This was an unexpected behaviour since in Canada, there are two official languages (French and English), and it is present in their daily *routine*, such as access to products and/or services in both official languages – product labels, medicines – official documents, radio, TV, and schools – students of elementary schools must accumulate 600 hours of French from grade 4 to 8 [Ont]. However, three of the English speaking participants mentioned they will use it when the *context* regarding location is relevant for sharing. Participants mentioned that they will use a foreign language to assign tags when moved by sharing contextual information, "*I use French when I travel through Canada.*" (P2), and "*I only use them in relation to the context.*" (P12), for instance.

At this stage we identified that the previous identification of user patterns of tagging could support recommender systems on the decision of recommending tags based on the context (location) users are in.

In both groups of participants from Brazil and Canada, there were participants who mentioned the use of foreign language when tagging makes you look "cool" or "fancy". Although this was not one of the main points mentioned by participants as reasons for tagging, it is something to take into consideration when looking for users' motivation for tagging in foreign languages.

²It is an idea, behavior, or style that spreads from person to person within a culture

Regarding propagation, English speakers think it is possible to accomplish it through the use of popular tags (pool of tags), that can help them reach a broad audience. This view will be discussed in detail in the next section.

4.2.2 Structure Choice

Regarding the structure of tagging, we found that when users assign syntagmatic tags, they are trying to share a *concept* or *context* of the image. These were the concepts we found from the analysis of tagging structure choice for **Portuguese speakers**.

At this point what became clear is that the structure of tags is strongly related to the wish to express image concepts and its context. Users can assign a combination of tags to describe a scenario/situation, the image context, a message that cannot be expressed by a single word. Moreover, syntagmatic tags are used to express an opinion, or even a “meme”.

Regarding the second question, during this stage, a participant reported:

P1: “(...) if an image has a set of things like a “campsite”, you know, a “tent”, a “campfire”, “people having fun”, so you may think this is maybe a barbecue (like a party). In my point of view, a tag should describe the “situation”, the scene, not just the objects one at a time.”

This participant uses composed tags and single words aiming to express - what we call Conceptual tags - the mental combination of photo characteristics, which results in tags, such as vacation, cold, party, etc. P3: “My personal insight is that a tag is an expression, not just a word.” For him, syntagmatic tags and tags that express a concept have the same function. It is indeed different to use tags that describe the objects on an image, from those that express the context of the image. Thus, the image context plays a role on the tags users choose to express themselves.

On the other hand, P5 expressed that she used tags to describe the content. P5: “(...) most of my tags are single words, sometimes I use more than a word to write a place’s name - if it is composed - maybe I can combine words with a verb or something else.” Differently from other participants, she uses tags for content indexing. This was a common behaviour among other participants during the study. As it was mentioned in the quantitative results of the previous study, context related images presented more syntagmatic tags, and paradigmatic tags were used to describe the image objects with the purpose of indexing.

English speakers think that more important than the language choice, is the tag structure used for tagging. They see syntagmatic tags as the main tool for propagation of content and its subject/context. P5: “I use [a tag with more than one word], because you can attract a wider range of people with the same topic.”

Many popular syntagmatic tags such as “followback”, “instamood”, have a meaning that those who are aware of their use will associate to theirs in an attempt to contextualize the image and associate it to other images (pool of images) that are from the same subject. For example, P12 is aware that some syntagmatic tags are trendy in the online community, which can increase his audience, at the same time as he uses them for contextualization. P12: “I use more than one tag to apply more context and link more trends.”

Participants also reported that syntagmatic tags also decrease the amount of messages users have to write to explain the context of the image. *P6: "I use tags with more than one word because it gives more detail to people than just using one word. This also reduces the amount of hashtags that need to be written."*

English speakers, similarly to Portuguese speakers, use tags as a way to simplify a message. *P2: "I use more than one word tags because some are statements or sayings."* This is an example of user who assigns tags that have a known meaning in the community, such as *"saturdayisfortheboys"*, *"friendzone"*. By using this pattern of tagging they are not only contextualizing the image but putting it in evidence to maybe collect more likes or followers.

4.3 Discussion

Based on the results we found at this stage in combination with the quantitative analysis we conducted on the data gathered during the experiment, we can conclude that the structure and language users choose for tagging is led by the audience and intention of indexing or contextualizing image content. At this stage we were able to address the differences among Portuguese and English speakers regarding their choices and motivation for tagging.

Adding personal insight to images was mentioned as one of the main reasons for tagging. The context of images described by tags reaches not only personal audience but could expand images to pool of tags increasing users audience by the language adopted in combination with the structure chosen for tagging.

For example, a user posting images online of an ocean view: the choice of tag structure will be guided by the audience he/she wants to reach and what he/she wants to communicate. The user could decide to describe the content of the image (for indexing content) or express his/her opinion, or the context of the moment (for contextualizing the image).

According to what we have found in our studies, when users want to send a "message" to the audience communicating the context of the image, they tend to adopt syntagmatic tags, such as *#summerbreak* *#Ideservelt* – the type of tag that reveals implicit details of the image. On the other hand, when the audience intended to be reached is beyond followers (personal audience), users may want to promote image content, not just the context, using tags that describe what actually is in the image (*#beach*, *#ocean*, *#peace*). There will be also cases in which users will assign syntagmatic tags to put images on a pool of tags or even for personal indexing (categorization). This is particularly interesting because users are indexing images through their context, and these differences are related to the frequency a tag is adopted.

Both groups of participants understand that tags are powerful tools to support search engines, so even when indexing is not their main goal for tagging, it occurs "naturally" in their tagging choice. Once they are motivated by the contextualization of the image, they will switch from one structure of tags to another. Participants made it clear that context sharing is in general the reason they use tags, and their choices are guided by a combination of personal insights, audience and goal/motivation

for tagging. They want to assign personalized terms, messages and words to describe the situation “taking place” at that moment. It is possible to identify that their behaviour switches according to their wish to express the context of the image (concept/situation), or indexing content (using tags to describe the image or put it on a pool of tags), as well as their target audience.

Furthermore, in addition to sharing personal tags, they try to reach distinct audiences through content (as well as their tags) they share using tags in different languages, or some popular/funny expression (e.g. “memes”). Assigning tags in other languages can reach distinct audiences without the need to write a long description of an image (which requires some knowledge of grammar in foreign languages) to express a situation, opinion, etc. Portuguese speakers feel safe writing tags in English, mainly using pool of tags/memes, when they know they will be understood by a broad audience. They may also use some “isolated words” or “common sense words” written in a foreign language to spread their image. It works as a tool for *indexing* content and expanding content audience. For this reason, tags are now used as a tool for communication by requiring just few words to express the image content and context.

Understanding how users perform the same task in different environments can provide insight for designers to decide among distinct approaches, according to user and system needs. With this goal in mind, we modeled the results we found in the previous studies to support the identification of tagging behaviour according to patterns of tagging. In the next Chapter, we present the steps we took to create this approach and accomplish the goals of this work.

5. A LANGUAGE-BASED APPROACH TO SUPPORT THE IDENTIFICATION OF TAGGING BEHAVIOUR

Tags are seen as a tool for communication, and even in those cases in which they are merely describing image content with paradigmatic tags, users have a target audience. Users want to convey a message (self-expression), and give hints about an image content (for searching (public) or for browsing (personal)), or to organize content (self-communication) [AN07, KKGS10, ZMS, SLR⁺06]. Based on the results we found during the user studies and the concepts identified by the qualitative analysis we have conducted, we concluded that users tend to switch the structure and the language for tagging, according to their motivation to do so, and the target audience intended to be reached.

As tags are textual data and sometimes their classification depends on manual analysis, which is impossible to be carried out when large amounts of data are available, we decided to design an approach to automatically identify tagging patterns and support the identification of tagging behaviour. We first present a model we designed to address the difference among tagging patterns combined with dimensions that represent users' motivation for tagging. This model is later used to create a framework to compute tags as quantitative data and support the identification of tagging behaviour. The concepts we found supported our awareness of who the tag readers are or can possibly be, and what is attempted to be expressed by users. Next we present a model to demonstrate how the patterns of tags and the motivation for tagging are related.

5.1 A Model of Tagging Patterns and Its Dimensions

One of the most important results we found in our work was the association of tagging patterns to users' motivation for tagging. Moreover, during the axial coding, we found communication to be a broad concept that is related to a subset of motivations that are linked to the audience and its role on users' decisions regarding language and structure.

As a result of our investigation, the model to be presented here has language and structure as patterns that change according to two main general motivations. Unlike other works, we are not interested in the identification of new or different motivations for tagging, since this topic was already investigated by other researchers and there is an extensive list of possible reasons associated to users tagging [GLYH10]. Instead, what we have noticed while discussing the results is that most of the motivations we have found in the literature could be summarized into two global motivations for tagging: contextualization and indexing. Table 5.1 shows the list of motivations reported by Gupta [GLYH10], and how we have organized it. In short, what we have found in our results is that although users have specific reasons for tagging, such as opinion expression, description of content, etc., they have shown two main task goals that are basically indexing or contextualization of images shared online.

Table 5.1: General motivations defined based on the subset of motivations already describe by the literature [GLYH10].

General Motivations	Subset of Motivations
Indexing	Future Retrieval Description, sharing and contribution To call attention Play and competition To earn money Task organization
Contextualization	Self-representation Opinion expression Contextual information

We will use these two general motivations against the structure users choose for tagging. By doing so we aim to identify how the patterns of tagging are related to users' specific motivations for tagging.

We have also noticed based on previous work [SKK10], that the frequency of tags could indicate that users are trying to categorize content. However, because of the nature of our previous experiment, we were not able to address these patterns in this stage of our model. For this reason, we will discuss these patterns during the step in which we create a framework (Section 5.2) to compute tags as quantitative data to support the identification of tagging behaviour.

In the next section, we will present patterns of tagging regarding its structure combined with the dimensions regarding users' motivation for tagging.

5.1.1 Structure

The concepts we found during the axial coding show that users assign tags not only motivated by content description or categorization, but also by sharing feelings, perceptions of the image context or concept, and expressing opinions and messages, all of which affecting their choices on which structure or language to use for such task. Mainly, the differences we found show the impact on tagging patterns used for self-expression [ZMS]. Users understand that tagging is a powerful tool for indexing content that can support search engines on information retrieval task, but the content description of an image was not reported as their main motivation for tagging.

Based on the combination of our findings and the list of motivations reported by Gupta [GLYH10], we narrowed our model down based on two general motivations – context and indexing – in which a subset of motivations is allocated according to users' choice of tagging structure. Table 5.2 shows these motivation differences and the choice of tagging structure. We display the structure of tags adopted in contrast to the motivation for tagging according to our findings [ZMS, ZBS16a], and the subset of motivations reported by Gupta [GLYH10]. This gives us a better perception on how the structure used for tagging is affected by users' motivation for assigning tags.

Table 5.2: Tagging structure versus the motivations for tagging.

		Motivation	
		Contextualization	Indexing
Structure	Syntagmatic	Self-expression Opinion expression; Contextual information;	Future Retrieval (categorization); To call attention; Pool of images; Personal tags.
	Paradigmatic	Conceptual tags; Mental combination of images characteristics;	Content over context; Description of content; Sharing and contribution.

These dimensions show how users have chosen a tagging structure according to their communication intentions, not only to contribute with the system goals regarding content description. Afterwards, we will discuss our model considering each dimension we have presented. Figure 5.1 shows the differences and relations among the chosen structures and their dimensions to support our discussion.

Syntagmatic & Contextualization

When motivated by sharing image context, syntagmatic tags are users' choice for tagging. Syntagmatic tags give more details about a subject, situation, or implicit content an image brings. These tags are likely to be shared to express opinion and contextual information. Figure 5.1 represents how tagging choices are related to their tagging motivations. Users who are motivated by self-expression have little interest in the indexing task. Their tags are there for that moment, to tell a story of the context. No putative tags score high for self-expression, once these tags are not generally from common sense use or very personal.

Syntagmatic & Indexing

These tags are assigned for personal indexing or putting personal content on a pool of images (public). The reason for that is that users tend to assign personal tags with the purpose of content organization. This kind of tags can bring context to the image at the same time as users are motivated by future retrieval. In Figure 5.1, we tried to represent the concept of indexing through the use of syntagmatic tags.

As we can observe from this situation, indexing and contextualization go hand in hand. This behaviour has organizational purposes as its main goal, which justify considering it as a category within the *indexing* motivation for tagging. They are used when users want to index an event, or a personal aspect of life. The tag `#Dave&AnasWedding` can be taken as an example of personal tags that are related to an event. These are syntagmatic tags that were clearly created with the intent of putting tags related to the same event in a pool of images. However, in order to identify this behaviour, we have to be able to identify the frequency of use of these tags. This is the difference

between contextualization and indexing motivations when using syntagmatic tags. These are aspects we will further address during the Case Study conducted as part of this work. In this case, users want to index images at the same time as they are concerned with the message they will convey.

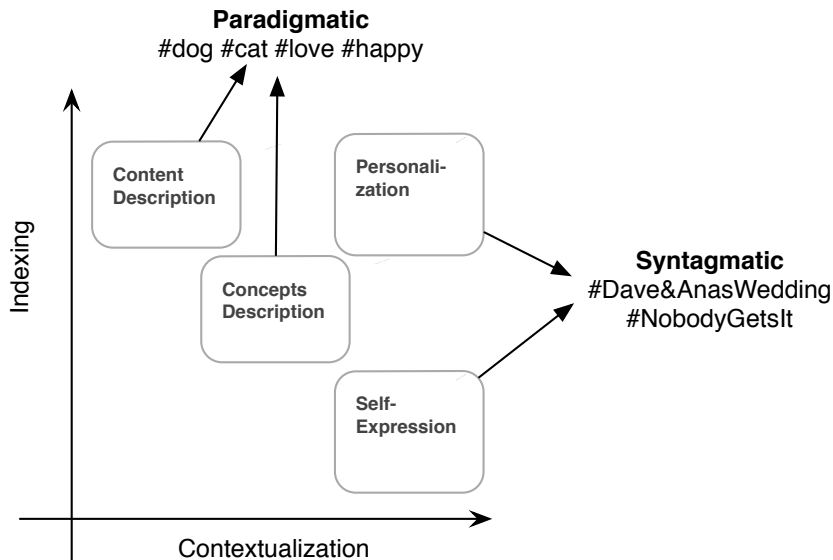


Figure 5.1: Indexing and Contextualization, motivation dimensions versus structure chosen for tagging.

Paradigmatic - Indexing & Contextualization

Although we have found in our studies that paradigmatic tags are mainly used for indexing content, the aspect of contextualization using paradigmatic tags has also been found as relevant for discussion, as reported in Table 5.2. We decided to explain both dimensions, indexing and contextualization, together to express the differences between them related to paradigmatic structure. Paradigmatic tags are generally used for content indexing over context. When users assign paradigmatic tags that are of common sense, it is reasonable to affirm that indexing comes over contextualization because of the use of putative tags. This is because these tags are used for content description, which puts images on the results of search engines that are related to a specific content. However, one point to observe is that although it is easier to index content using paradigmatic tags, contextualization of images can also be expressed. For example, concepts such as *#love* or *#happiness* are not objects present in an image, but else the mental combinations of the elements present in an image that create concepts [ZMS]. We try to better represent how motivations for tagging can mediate users' choices for tagging in Figure 5.1. Concept description is in the middle of indexing and contextualization motivations because of the nature of the word (adjectives – happy, sad). Although not covered by the scope of this work, the use of semantic analysis to identify paradigmatic tags with the purpose of contextualization could support designers to understand if users want to index content or contextual information. On the other hand, paradigmatic tags are in general used for content description, and have indexing as main motivation for their use.

5.1.2 Language

The language used for publishing content online has implications for the dissemination of content [RGH⁺14]. Assigning tags in other languages can reach distinct audiences without the need of writing a long description of an image (which requires some knowledge about grammar and vocabulary) to index or contextualize its resource subject matter or the content it brings. As a result of this behaviour, non-native English speakers, for example, may only use some “isolated words” and/or “common sense tags” written in a foreign language to spread an image. For this reason, tags are now being used as a communication tool among cultures by requiring just a few words to express image content and context.

During the studies we have conducted, Portuguese speakers demonstrated concern regarding who was “listening” to their messages. They showed self-consciousness perceiving they would reach more people when using a foreign language than assigning tags only in Portuguese. On the other hand, English speakers did not express any concern regarding who would be “listening” to their tags on the matter of the adopted language. Table 5.3 shows the differences between English and Portuguese speakers. The main motivation we found for English speakers to assign tags in a foreign language was grounded on the image context regarding location. Other languages were used for tagging when context localization was involved, e.g. “*tour eiffel*” for the Eiffel Tour. This tag was used because the recommender system suggested it. Differently from Portuguese speakers, English speakers feel that everybody is listening to them. Instead of the language used, they see the structure of tags as a resource to expand their audience.

Portuguese speakers use putative tags in a foreign language mainly to describe content (paradigmatic tags), which includes location, as English speakers do. They can also assign popular tags to expand their audience (statements, memes, syntagmatic tags). It is easier for Portuguese speakers than English speakers to use statements in foreign languages when they want to express the context, because many of them are aware of the tag context, e.g. #ThrowbackThursday¹. In Figure 5.3, we represent a trend found among Portuguese speakers. They are more prone to use paradigmatic tags than syntagmatic tags. They use tags in English not only to index or contextualize an image, but to expand their audience. Although the language domain is a concern, they are exposed to many popular tags online due to social media resources, such as trendy topics on Twitter and Facebook that show popular hashtags and memes indexing content. Popular hashtags can be used to expand their audience (indexing), and express the image context (for contextualization of the subject matter). For example, a pool of tags, such as “*like4like*”, sends a message to other users that *if you give a like on my image I will do the same to yours*. However, English speakers do not see foreign language tags as an everyday tool for spreading content. They mentioned using foreign languages to assign tags when there is context location as evidence in the image.

For both English and Portuguese speakers, however, their choices will be guided by their general knowledge of the language. It indicates that putative tags are generally the type of tags that users

¹Nostalgic personal images from the past that are posted on Thursday, followed by the #ThrowbackThursday tag

Table 5.3: Indexing and Language Domain were two main factors found to lead users to assign or not tags in foreign languages.

Multilanguage Motivation	
English speakers	Portuguese Speakers
Location-related	Content Indexing (description) Context Indexing (pool of tags, memes)

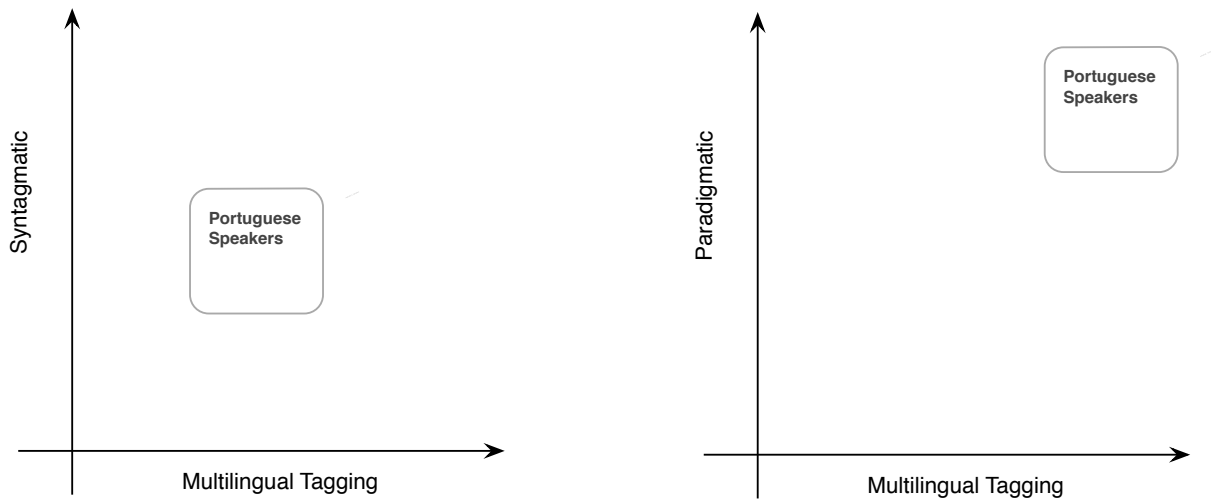


Figure 5.2: Tendency among Portuguese speakers: assigning putative tags (paradigmatic structure, in general), popular tags (memes, pool of images related to tags – #ThrowbackThursday), or accepting tag recommendation.

will choose to assign. They do not want to risk writing messages in foreign languages due to their concern about language domain. Despite that, they will assign tags in foreign languages that are of common-agreement, popular, or recommended by a system, if they know their meaning.

5.2 A Framework for the Identification of Tagging Behaviour

At this stage we present a framework to compute tagging patterns as quantitative variables² in order to support the identification of tagging behaviour, and replicate the model we designed. Through this framework we expect to compute the most adopted tag structure, the most used language, and other features, such as repetition of tags, variability and so on, by each user in a real world dataset. These features, alongside clustering tools, will support the identification of tagging behaviour on tagging datasets from Instagram and Flickr.

In order to create this framework and the features based on tagging patterns, we need to be able to automatically identify the preferred language and the structure used for tagging from the setting of tags of each user's profile. Our aim is to be able to compute users' tagging patterns as integer numbers later used to compute features that represent tagging behaviour for a user.

²Variables represent integer number later used to compute the features.

A user's profile used for the purpose of this framework follows the same structure adopted on traditional tagging system, which consists of a triple composed by a user id, an image id, and a set of tags. It means that each user u has at least one tag t or a set of tags T associated to an image/resource r , which is the same as assuming that it is a post $P = \langle u, r, T \rangle$ composed by these three elements. Following this structure, we decided to compute features that could indicate the main structure and language adopted to each profile available on real datasets of tags. Hence, at the end, each user will have features that represent his/her tagging behaviour.

The first challenge we found in order to create this framework was to be able to find out the language for those cases in which tags had a syntagmatic structure or when they were assigned using other languages apart from Portuguese or English. For this reason, in the case study to be reported in the next Chapter 6, we worked with a dataset whose gathered resources were from users assigning tags in which the metadata regarding the image location were from countries where English and Portuguese are more likely to be spoken. Also, there will be cases in which it will not be possible to identify or categorize tags as syntagmatic or paradigmatic, because some words, such as proper names, slang, acronyms, and so on, are not present in the dictionaries we use. Therefore, unknown tags will also be computed, and we aim to use them as a source to differ users' behaviour.

We intend to identify the number of tags each user assigned in English or Portuguese, and how many of them have syntagmatic or paradigmatic structure. Moreover, we also seek to compute the following features by each user:

- Frequency of repeated tags;
- Frequency of repeated tags in English/Portuguese with paradigmatic/syntagmatic structure;
- Heterogeneity of tags, from the set of repeated tags.

Figure 5.3 shows an overview of the process of computing tagging features that consist of two major steps, namely language identification (Section 5.2.1), and tag segmentation of syntagmatic tags (Section 5.2.2). In this example, the tag `#bestFriend` is not recognized in the language identification stage because it is not present in dictionaries available for comparison. Thus, it goes to the next stage, where tag segmentation occurs. First, it will go to the English Corpus if the user-data was gathered from the English-speaker location defined by the API we used. After processing language identification, tag segmentation, and computing those tags that are repeated, each profile will have resulting variables that will later be computed as features. In order to accomplish this task, we resorted to a combination of tagging segmentation process using a word segmentation library, a generic spell checking library, and two corpora in English and Portuguese. After that, we formalized the steps we took in order to compute the features that will be used in the process of tagging behaviour identification. Next we will explain the used tools to identify tagging patterns and how they were computed.

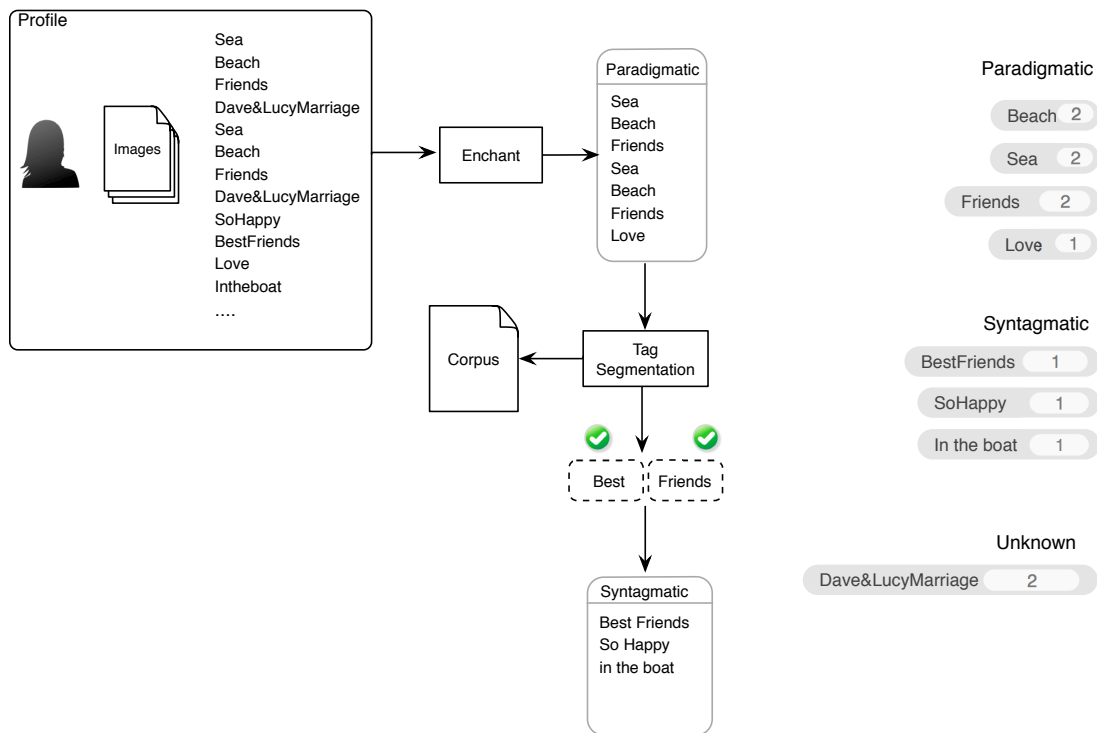


Figure 5.3: All tags go to the first step of language identification through a library called Enchant. The tag #bestFriends went through the process of Tag segmentation since it was not identified as a word present in Portuguese and English dictionaries.

5.2.1 Language Identification

In order to identify if a tag has syntagmatic or paradigmatic structure, we first need to analyze if a tag is available in a dictionary or not. Figure 5.4 shows an overview of how we compute tags according to their language and structure. For example, we can easily find the word “beach” in an English dictionary, but we will not find the tag “friendsfromcollege” in it, because this tag consists of a string without any element that says it brings more than one word together.

As we implemented the framework, we had to resort to the use a generic spell checking library to support us on the identification of tags as a unit – paradigmatic tags. The Enchant [Lac17] library consists of an interface that provides a comprehensive way to work with different types of spell-checker libraries, such as Hunspell (formerly Myspell), GNU Aspell, Hspell, Apple Spell (macOS only), among others. We implemented our framework using Hunspell [hun] spell-checker library and adapted the PyEnchant package [Kel] to support the identification of tags in Portuguese. PyEnchant is a python module that uses Enchant spell checking library. This module allows adding dictionaries other than English. Originally, the PyEnchant module is able to identify the following languages: British English, American English, German, and French. Hunspell dictionary, used by LibreOffice, was used as source for spell checking words in Portuguese. All resources used for this stage, such as Enchant library, PyEnchant module, and Hunspell dictionary, are available under the General Public

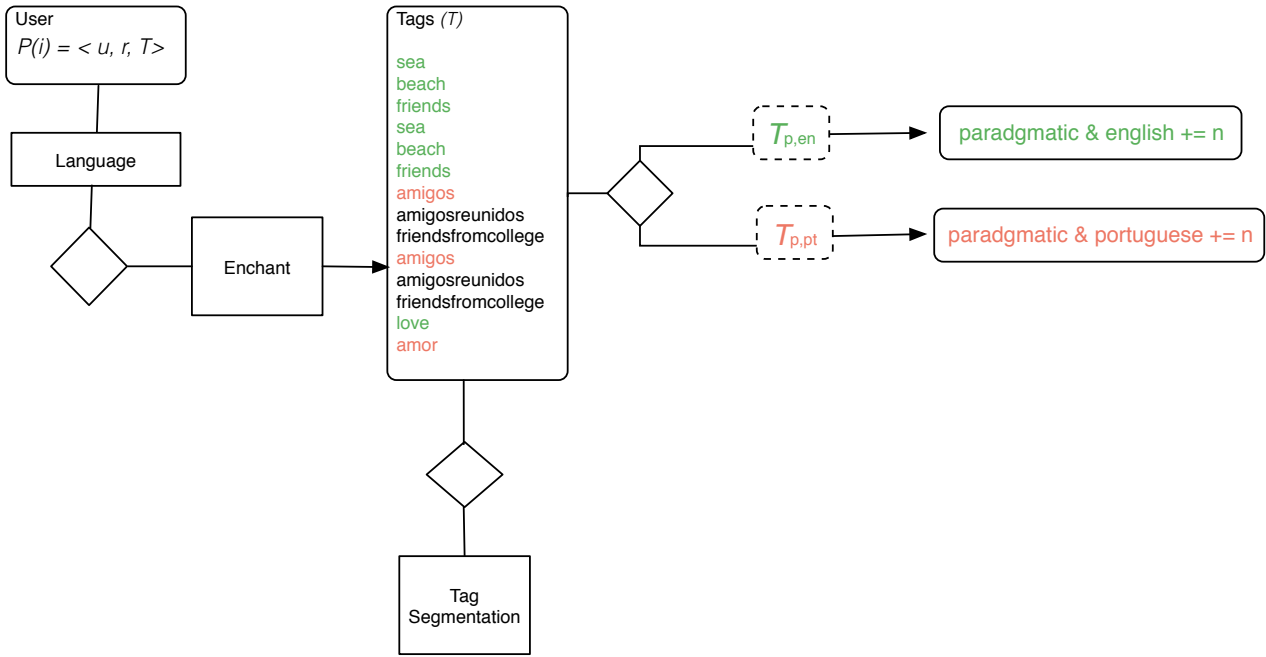


Figure 5.4: Language identification steps during the identification process of paradigmatic tags.

License (GPL). As we know that most tags in a dictionary represent a paradigmatic structure, we could count paradigmatic tags based on the output from the spell checking library.

The process consists of the analysis of tags for each user. Based on the previous formalization for postings, for each u there will be many posting P which are defined by a triple. One user can have many different triples associated to him/her $P_i(u)$, but each triple is associated to only one user. However, the tags t that are part of each set of tags T can repeat in different triples. As an example, shown in Figure 5.5, a user has three different posts P , each of them with three different images. To each image there are three different sets of tags. There will be repeated tags in each set, which can be represented by the intersection of sets of tags, an issue that will be later discussed in depth.

Once a paradigmatic tag is identified as being part of a dictionary, we shall compute the following step:

$$isParadigmatic(t, D(l)) = \begin{cases} 1, t \in D \\ 0, t \notin D \end{cases} \quad (5.1)$$

where t is a tag and D is a dictionary for a target language l . Followed by the function $paradgmatic(t, l)$ that computes for all tags of a user u how many of them exist in the set of words of a dictionary D for a given language l :

$$paradgmatic(t, l) = \sum_{P_i \in P(u)} isParadigmatic(t, D(l)) \quad (5.2)$$

This function runs over all tags of one user at a time. Every time one tag is identified as a word that exists in the dictionary, the function will sum it up to the user's related feature. For



Figure 5.5: Example of a user that has three posting P . Each posting is composed by an image (r_i), and set of tags T in which $T = t_1, t_2, \dots, t_n$.

example, the tag “friends” would be considered a paradigmatic tag because it would be recognized as a word that is present in the English dictionary. However, if a tag is not recognized by any of the dictionaries, it will go to the next step, which consists of the tag segmentation process. The resulting variables and features based on the results of the function will be further detailed.

5.2.2 Tag Segmentation

One of the most important outcomes we have had was related to the association of tagging structure and user motivations for doing so. We seek to identify these differences not only by identifying paradigmatic tags, but by considering all residual tags for analysis instead of just assume it as syntagmatic. Otherwise, it would completely put away the language being used for syntagmatic tagging, which would go against the model we have created. In order to identify the structure and language of each tag, we need to go through two steps for each target language used in the datasets we will analyze. Following the previous approach to identify paradigmatic tags, each tag that was not identified as paradigmatic will go through the steps we will present next.

Take into account the following set of tags $T = \langle \text{friendsforever, amigosparasempre, amigos, friends} \rangle$. What we notice according to our model, by looking at such data, is that this user is expressing the image context, as well as the content it has got (people who are friends). Moreover, this user

adopted a multilingual tagging behaviour, and this is only possible to be identified because we know that Portuguese and English are being used. However, in order to compute that “*friendsforever*” and “*amigosparasempre*”³ are syntagmatic tags in English and Portuguese, respectively, we need to be able to split these tags into words. This approach is called word segmentation [SH09] and it is used by search engines to split, for example, URLs into words, and try to identify what they are about. For instance, the output of the word segmentation process of URLs “*sportcheck.com*” and “*dollartree.com*” would result into “*sport check*” and “*dollar tree*”. In combination with the previous step of language identification, we used a Word Segmentation python library [Jen] that counts with an English Corpus from Google that contains one trillion words and a subset of 330,000 unigrams. In order to use the same library to segment tags assigned in Portuguese, we used a Portuguese Corpus from CETENFolha [SB00], that has a collection of approximately 24 million words in Brazilian Portuguese and 241,392 unigrams. The words were collected from an online newspaper called Folha de São Paulo.

Figure 5.6 shows how the library splits strings of text into single words. This approach uses the support of a Corpus in a given language to compute the probability of a word to be part of a string of words in which there is no space to delimit what a meaningful word is or is not.

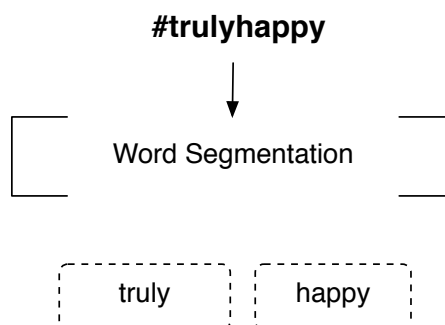


Figure 5.6: Word segmentation approach used for the identification of tags that have syntagmatic structure.

However, as many other approaches that aim at text identification, the word segmentation algorithm also has some drawbacks. For example, in Figure 5.7, we can see that even if a word is not a real word, the algorithm will put apart the words that are not really meaningful as single words as an attempt to identify all possible valid units in a string. We then included one more step in the stage of tag segmentation that consists of inspecting if a string is a meaningful word. Through the support of the same spell checker used during the previous stage, we inspected each word split by the word segmentation algorithm and verified if they appear in the dictionary in the language being analyzed. Then, if the quantity of words found for a tag by the word segmentation algorithm was the same found as real words in a dictionary, we would consider them as syntagmatic tags. For example, a tag that is considered syntagmatic is indeed a set of sub tags $T_s = \langle \textit{love, you, so, much} \rangle$, and then we first verify if each word in the set is present in the Corpus C for Portuguese or English,

³The tag “*amigosparasempre*” is the literal translation in Portuguese of the tag “*friendsforever*” in English.

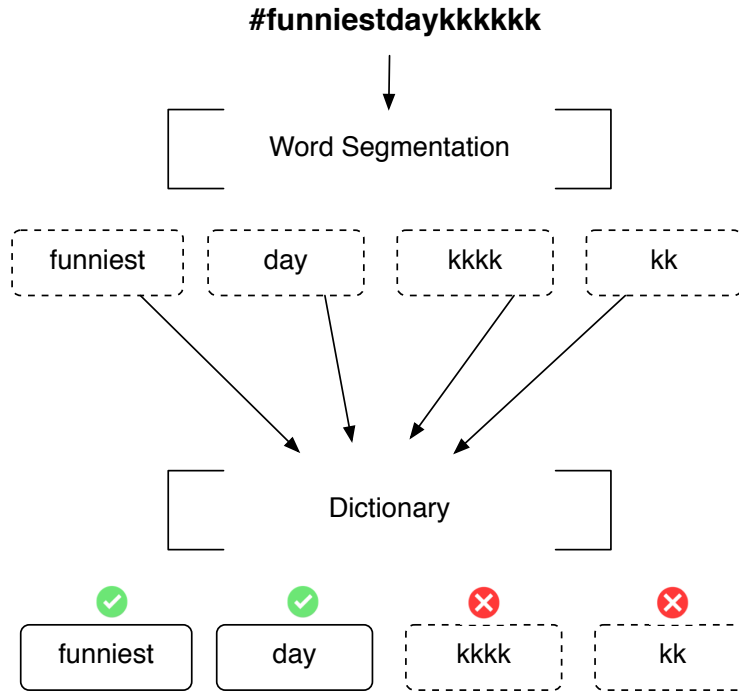


Figure 5.7: Even when words have no real meaning, the algorithm will try to split them.

where l is the language (Function 5.3). Every word that is part of a syntagmatic structure should also be present in the Corpus and in the dictionary for the target language.

$$isSyntagmatic(t, C(l), D(l)) = \begin{cases} 1, \forall t \in T_s/T_s \subset C \wedge t \in D \\ 0, \exists t \in T_s/T_s \not\subset C \vee t \notin D \end{cases} \quad (5.3)$$

Then, Functions 5.4 are used with the intent of computing the final result of syntagmatic tags' frequency for a user profile by

$$syntagmatic(t, l) = \sum_{P_i \in P(u)} isSyntagmatic(t, D(l)) \quad (5.4)$$

As a final step, if a tag was not found as being part of the set of syntagmatic or paradigmatic tags, we then computed it as an unknown structure, using the same Functions but without the language parameter:

$$isUnknown(t) = \begin{cases} 1, \exists t \in T_s/T_s \not\subset C \vee t \notin D \\ 0, \forall t \in T_s/T_s \subset C \wedge t \in D \end{cases} \quad (5.5)$$

We then used Function 5.6 intending to compute the final result of unknown tags' frequency for the user profile by

$$unknown(t) = \sum_{P_i \in P(u)} isUnknown(t) \quad (5.6)$$

5.2.3 Repetition and Heterogeneity of tags

Repetition could indicate that users are indexing content by using the same tag repeatedly. The analysis of tag use frequency regarding its structure could give us insights on what kind of content users are trying to index and their motivation to do so. We decided to conduct this step as an additional resource to support the task of identifying tagging behaviour. We computed the quantity of repeated tags to each feature we have found as support to identify indexing or contextualization behaviour alongside with the language and structure of adopted tags.

Because we seek to be able to know which structure is being repeated, the following steps are applied to each set of tags found during the stages in which we computed paradigmatic and syntagmatic tags' frequency. It means that, besides computing the repetition for all sets of tags from a target user, we also computed the set of syntagmatic, paradigmatic, and unknown tags that occurred more than once for that user. To do so, we computed the k -tags that have occurred more than once in the user's set of tags. We used the following function to address the matter:

$$isRepeated(t, u) = \begin{cases} 1, & |t| > 1 \\ 0, & |t| \leq 1 \end{cases} \quad (5.7)$$

This function will signalize if a target tag occurred more than once in the set of tags for the target user. The total number of repeated tags can be compute by:

$$repetition(t, u) = \sum_{P_i \in P(u)} isRepeated(t, u) \quad (5.8)$$

We also want to be able to identify how the heterogeneity of repeated tags could be related to users' tagging behaviour. As we can see in Figure 5.8, there is a difference among the sets of tags of user u_1 , who has many different tags being used and repeated, compared to user u_2 , who always repeats the same tag.

To compute the heterogeneity of repeated tags for a given user, we compute the number of tags that are different in the set of repeated tags. In order to do so, we infer the intersection of tags from each image by computing the amount of different tags that has been repeated at least once by

$$heterogeneity(t, u) = \sum_{P_i \in P(u)} |t_i \cap t_j| \quad (5.9)$$

These same functions 5.8 and 5.9 are used to compute repetition and heterogeneity of repeated tags in the set of tags classified as paradigmatic, syntagmatics or unknown.

Figure 5.9 shows how we approach computing the heterogeneity of repeated tags. We believe that measuring tagging repetition and heterogeneity could support in the task of clustering analysis by differing users among those that are motivated by indexing or contextualization of content alongside the tagging patterns we have modelled.

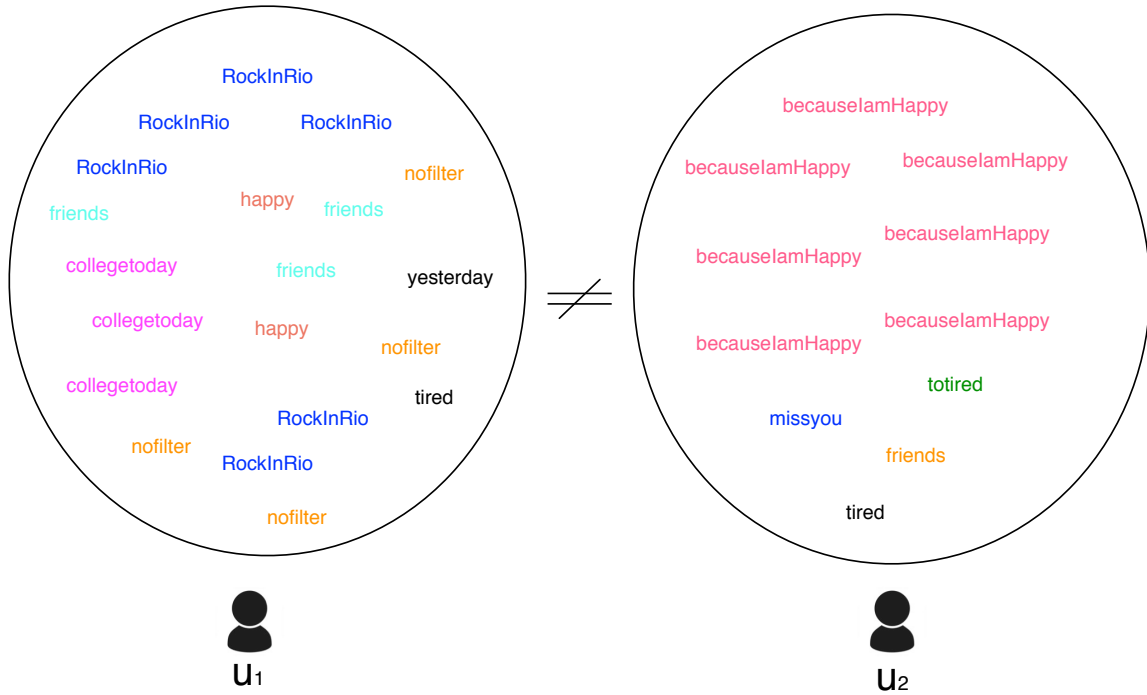


Figure 5.8: Difference in use of repeated tags.

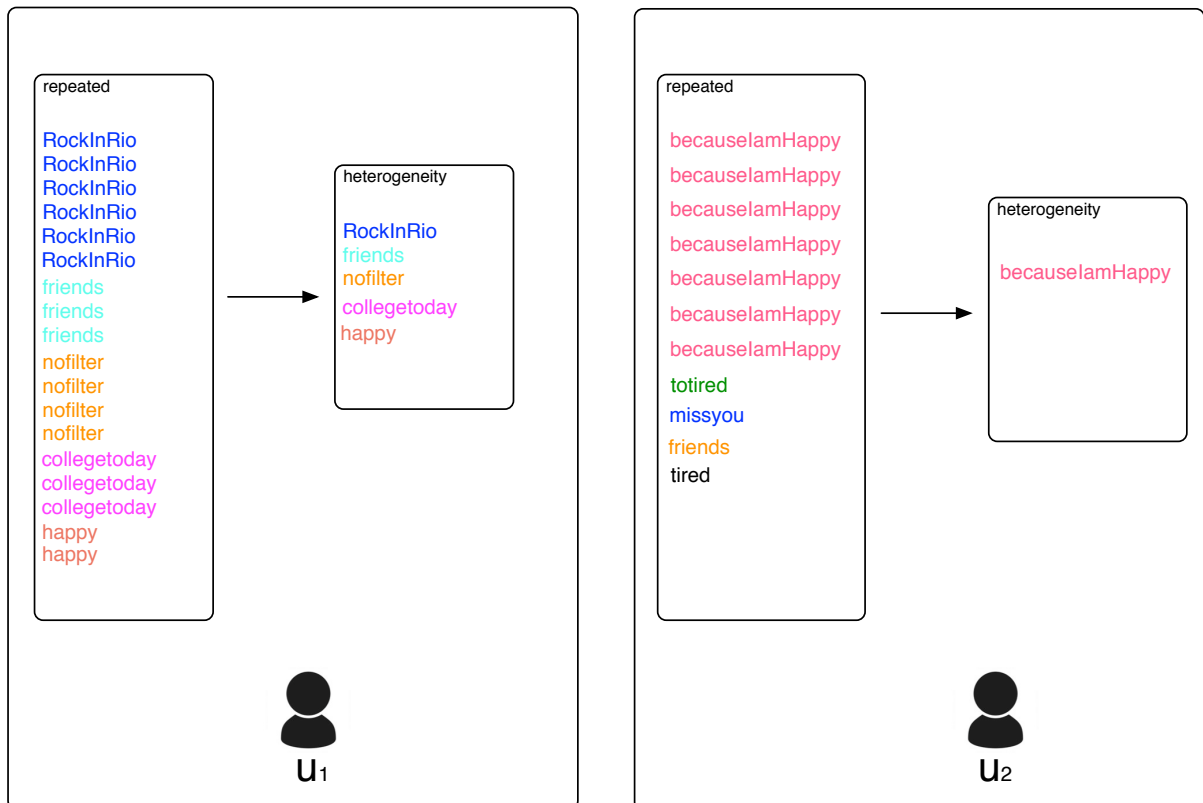


Figure 5.9: Heterogeneity of tags and the difference in the use of tags as resource for content indexing.

After computing all variables, we are able to compute them as features, that is, a meaningful number that gives us insight about individual users' tagging behaviour. We use the functions

presented in the last section to compute the features presented in Table 5.4 as proportions. The reason for choosing proportion as a final value for our features is because we intend to compare profiles not from the amount of tags they have used, but on how the use of tags regarding structure, language, and repetition could support the identification of tagging behaviour for one user at a time. For example, instead of getting the total number of syntagmatic tags to a target user u_i , we compute the proportion it represents to the profile. This could be addressed by getting the total of syntagmatic tags we found $syntagmatic(t, l)$ to a given user, divided by the total number of tags $|t|$ assigned by the same user. Table 5.5 shows an overview of the results of each row to each feature by each user. By computing features for each user in a dataset, we will provide a matrix as a result, in which the rows are represented by users' ids and the columns by the features we computed to each user following the functions we presented and their proportions for each profile. We intend to be able to look at each profile and have a general understanding of his/her tagging behaviour by the combination of variables computed to each user profile. Moreover, these features will be the data we will use to cluster users by similar tagging behaviour during our Case study.

However, one of the biggest challenges of quantitative data is that it depends on the designer's ability to identify which features are relevant to capture users' behaviour, their needs, motivations, and so on. To overcome this issue we can resort to the model we have designed to make assumptions about users' intentions when tagging. Based on the findings of each study we conducted and the background from the field of tagging, we were able to generate 12 general features that can be expanded to 19 when Portuguese and English are likely involved in the dataset vocabulary to address users' differences on the adopted language, for example. We expect to analyze how patterns are associated in the Case Study following the Framework results, in which we use a real dataset to identify patterns of tagging behaviour.

Table 5.4: Framework features to support the identification of tagging behaviour.

Feature	Description	Computation
[eng/pt]__parad	Proportion of paradigmatic tags.	$paradigmatic(t, l) / t $
[eng/pt]__rep__parad	Proportion of paradigmatic repeated tags.	$repetitionParadigmatic(t, u) / t $
[eng/pt]__het__rep__parad	Heterogeneity in the set of paradigmatic repeated tags, also computed as proportion.	$heterogeneityParadigmatic(t, u) / repetitionParadigmatic(t, u)$
[eng/pt]__synt	Proportion of syntagmatic tags.	$syntagmatic(t, l) / t $
[eng/pt]__rep__synt	Proportion of syntagmatic repeated tags.	$repetitionSyntagmatic(t, u) / t $
[eng/pt]__het__rep__synt	Heterogeneity in the set of syntagmatic repeated tags, also computed as proportion.	$heterogeneitySyntagmatic(t, u) / repetitionSyntagmatic(t, u)$
[eng/pt]__prop	Proportion of paradigmatic according to the language being investigated.	$(syntagmatic(t, l) + paradigmatic(t, l)) / t $
total__repeated	Proportion of repeated tags.	$repetition(t, u) / t $
total__repeated__het	Heterogeneity in the set of repeated tags.	$heterogeneity(t, u) / repetition(t, u)$
unknown	Proportion of unknown tags.	$Unknown(t, l) / t $
unknown__rep	Proportion of unknown repeated tags.	$repetitionUnknown(t, u) / t $
unknown__het__rep	Heterogeneity of unknown repeated tags, also computed as proportion.	$heterogeneityUnknown(t, u) / repetitionUnknown(t, u)$

Table 5.5: Overview of the features as a final result computed by the framework.

id	eng_parad	eng_rep_parad	eng_het_rep_parad	eng_synt	eng_rep_synt	eng_het_rep_synt	...
01	0.32	0.0	0.0	0.35	0.20	0.50	...
02	0.52	0.57	0.24
...

6. CASE STUDY

This chapter describes the results obtained during a case study we conducted in order to replicate our approach on data gathered from real world tagging systems. To accomplish this task we used the framework we previously described to compute tags as features from two datasets, Flickr and Instagram, with the intent of identifying tagging behaviour through a clustering tool. We conducted the case study using two different datasets in order to verify the framework outcomes in different tagging environments. We start this Chapter presenting the datasets we used, followed by the framework application. After that, we introduce the chosen clustering tool in order to identify groups of users that have similar tagging behaviour. We discuss insights and shortcomings in the findings section and at the end we resort to Personas as a tool for discussing the clustering results.

6.1 Data Collection, Data Pre-processing, and Framework Application

The volume of data available online makes it impossible for designers to conduct tagging behaviour identification exclusively through manual analysis. Following the steps of our framework, one can transform raw tags into quantitative data to be used as source to produce clusters and identify tagging behaviour without demanding expertise in the field. To conduct this case study we collected tags from Instagram and Flickr social media networks. Instagram and Flickr provide APIs that allow gathering users' public data available online. We collected tags to compute features to use them on cluster tools and verify whether it is possible to identify groups of users that share the same tagging behaviour, and how the model we designed applies to real world data from tagging applications. Also, we expect to be able to identify aspects such as repetition of the same set of tags as evidence of personal indexing, since we were not able to address this behaviour by the user studies we have conducted previously.

6.1.1 Instagram

In order to narrow our analysis mainly to images assigned with tags in English and Portuguese, we decided to collect data presetting the localization where users are more likely to use these two languages. This action resulted in two different datasets, both from Instagram, one whose tags were assigned mainly in English, from USA and Canada, and another with tags from Brazil, assigned in Portuguese and English.

Collecting data assuming users' geolocation has been used for assuming users' cultural background [GGQJ13]. Instagram API allows us to search for recent media in a given area based on latitude and longitude, and collect content information about users that have posted images in some specific location.

For example, if an Instagram user posted a public image while at the Sugar Loaf¹, and has added this information to the image by geo tagging it, it is possible to collect data regarding the amount of “likes” it has received, the tags assigned, the comments it has got, the user’s id for the image, and so on. In order to access tags assigned in Portuguese and English, we narrowed our data collection by targeting the latitude and longitude from state capitals from Brazil, USA, and English speaking provinces in Canada. After setting the locations, we tried to identify if a given image randomly collected had tags assigned on it. We then collected the Instagram user *id* and looked for the collection of recent images this user had publicly posted online, requesting a maximum of 120 images from each user. Unfortunately, the rate limit for data collection is controlled by the time and number of requests. So, each time we reached the limit, we had to manually attribute a new location for an hour window of requests. In total we collected 535,006 tags from USA and Canada from 1,382 users, and 350,406 tags from Brazil from 944 users.

We computed the framework features to each one of the datasets. After applying the framework, we were able to analyze each feature and decide among the resulting data, which would be useful to conduct the clustering analysis. For example, due to the fact we found in our previous studies that in Brazil people are likely to use tags in English, we decided to compute the structure of tags in English used in the dataset from Brazil. Table 6.1 shows the summary of the data we have collected from Instagram. From the set of tags collected from USA and Canada, 77% (413,684) of them were positively identified by our framework as tags in English – paradigmatics + syntagmatics. From the set of tags collected from Brazil, 28% (100,615) were identified as tags in Portuguese, and 35% (125,391) in English – paradigmatics + syntagmatics. The other portion of tags could not be identified as tags in English or Portuguese, so they fell into the class of unknown tags.

Regarding the dataset in English, the framework computed an overall of 167,668 paradigmatic and 246,016 syntagmatic tags in English, besides 118,476 tags that were not recognized by the framework. We did the same data observation on the dataset collected from Brazil as shown in Table 6.1.

At this stage after collecting these variables, we computed them as features using the functions we mentioned in the previous Chapter, meaning that they will result in proportions for being used in the step of clustering analysis. Moreover, we conducted a data cleaning step, identifying outliers in order to cut them off and decrease data sparsity. We performed the cleaning step by visualizing the dataset amount of tags and the average of tags assigned by each user. Figure 6.1 represents the dataset of English Speakers on Instagram and Figure 6.2 shows the Portuguese speakers’ amount of tags.

We narrowed the minimum number of tags by users to 20 and the maximum to 4,000 tags in order to cut off data that could represent very low or high results when computing features. This action reduced the amount of data we used during the clustering task and the sparsity of data caused by users who had too many or too few assigned tags. As a result, the number of users, tags, and images were reduced, resulting in an English dataset with a total of 1,159 users, 429,419 tags,

¹Touristic point in Rio de Janeiro, Brazil.

Table 6.1: Instagram dataset and summary of framework results.

	US & Canada	Brazil
Users	1,382	944
Images	151,843	104,068
Tags	532,160	350,406
Parad. EN	167,668	64,610
Parad. EN Repeated	123,963	53,509
Synt. EN	246,016	60,781
Synt. EN Repeated	168,453	47,301
Parad. PT	–	41,718
Parad. PT Repeated	–	29,996
Synt. PT	–	58,897
Synt. PT Repeated	–	40,353
Repeated	379,468	262,560
Unknow	118,476	124,440
Unknow Repeated	87,052	91,401

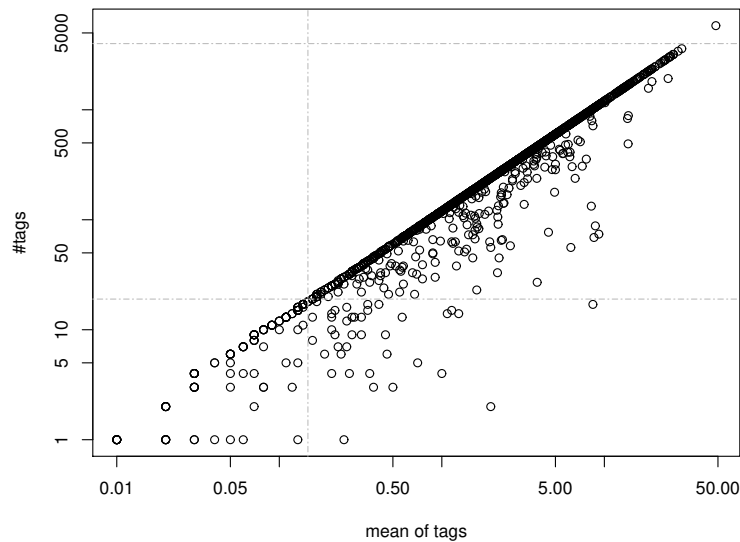


Figure 6.1: Cut off of users with less than 20 tags assigned in total – Instagram dataset, English.

and 131,435 images, while the Portuguese dataset resulted in 799 users, 343,616 tags, and 91,037 images (Table 6.2).

6.1.2 Flickr

With the intent of comparing how our framework would support the identification of tagging behaviour on systems that have different goals, we decided to use it with data from Flickr dataset. Differently from Instagram dataset, data gathered from Flickr were not filtered based on geo location or even assuming the language used by the users. This is the same dataset used as training data for

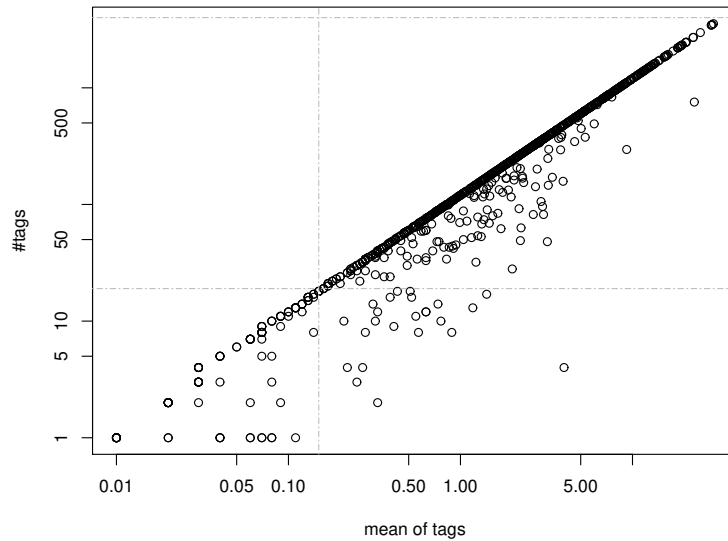


Figure 6.2: Cut off of users with less than 20 tags assigned in total – Instagram dataset, Portuguese.

the recommendation approach we presented in the user studies. The only condition we have defined while collecting data was that each image should have had at least 2-6 tags assigned to help the recommender algorithm perform relevant recommendations [dCZdO13]. However, in order to reduce the amount of data for computing the features, we tried to filter the languages used in order to narrow our analysis to Portuguese and English tags. We sorted out users who had images classified as English and/or Portuguese, and "others"². The original dataset had 605,403 tags and, after preprocessing the data using the same language classifier from the user study [LB11], we ended up with a new dataset that had 568,433 tags and 36,382 users. This means the preprocessing stage has estimated that each user had at least one tag assigned in English or Portuguese. When we applied the framework, it indicated that 471,256 tags were assigned in English and 17,655 were assigned in Portuguese. The other tags were either assigned in other languages or could not be recognized by the framework.

Although the dataset had a representative number of tags to analyze, we conducted the same step we did on Instagram to reduce data sparsity. However, because this dataset was gathered using a different requirement of data selection, we found that observing the number of images would be more relevant than the number of tags assigned in the cut off step, since the minimum and maximum number of tags assigned is known and delimited during the process of data gathering. Figure 6.3 shows the number of images by the mean of tags assigned by each user in the dataset. What we see is that it has an expressive number of users that have less than 20 images associated to their profiles. Similarly to what we did during the cut off of Instagram dataset, we also kept 20 as the minimum number of images a user should have had in order to use their tags as source to

²Assuming that these cases could represent syntagmatic structures that were not at first recognized by the language identification algorithm.

Table 6.2: Instagram dataset resulted after the cut off of *ids* that had less than 20 assigned tags. Our goal is to decrease data sparsity and improve results of the clustering analysis we will conduct.

Resulting Variables	Canada & USA	Brazil
Users	1,159	799
Images	131,435	91,037
Tags	426,970	343,616
Parad EN.	132,881	64,137
Parad EN. Repeated	92,773	53,146
Synt EN.	198,976	60,399
Synt EN. Repeated	126,480	47,099
Parad PT.	–	41,513
Parad PT. Repeated	–	29,933
Synt PT.	–	58,433
Synt PT. Repeated	–	40,097
Repeated	285,425	256,952
Unknown	95,113	119,134
Unknown Repeated	66,172	86,677

the clustering task. We also reduced to 450 the maximum number of images by user to keep out of the threshold those visible users that behave as outliers, as we can observe in Figure 6.3.

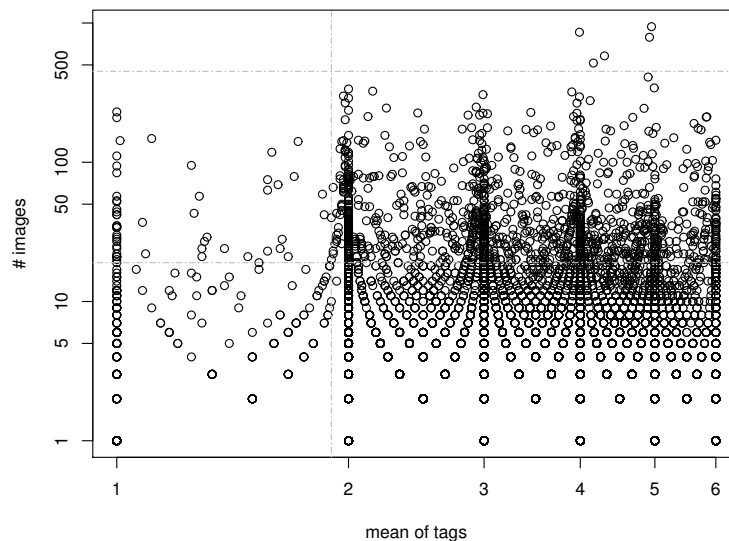


Figure 6.3: Flickr dataset and its data cut off to reduce sparsity and improve the quality of results in the clustering task.

The cut off narrowed the dataset to 1,261 users, less than 4% of users from the initial sample. Anyhow, this small representation of users hold more than 60% of the tags we have originally collected. Table 6.3 shows the final data summary we kept after data cleaning. They are presented as a summarized results' table to represent the final data collected for the clustering task.

Table 6.3: Flickr dataset, before and after data cut off.

Features	Before	After
Users	36,382	1,261
Images	155,251	64,061
Tags	568,433	343,616
Parad EN.	385,020	140,853
Parad EN. Repeated	276,774	137,598
Synt EN.	86,236	39,535
Synt EN. Repeated	69,697	38,735
Parad PT.	13,691	7,243
Parad PT. Repeated	11,458	7,214
Synt PT.	3,964	1,983
Synt PT. Repeated	3,363	1,950
Repeated	422,824	220,523
Unknown	79,522	36,192
Unknown Repeated	61,532	35,026

Next, we will present the steps we conducted to the clustering analysis and how we decided among the best algorithm for clustering the data we have computed.

6.2 Clustering

One of the challenges designers face while building an application is to be able to carry out the task of identifying patterns of behaviour to select data requirements across different types of context and subjective data [MLFA11]. Clustering is an unsupervised learning technique that supports the identification of patterns of behaviour while exploring data [Lay15]. From the best of our knowledge, no other work has used tags as quantitative data combined with cluster algorithms for the identification of tagging behaviour. By using the framework we presented in the previous Chapter, we transformed raw tags gathered from Flickr and Instagram into quantitative information in order to use it on clustering algorithms. By doing so, we seek to identify not individual tagging behaviour, but tagging behaviour similarities among groups of users.

By clustering users through the combination of the features we created for the framework and comparing them against our model, we expect to identify different sets of users that can contribute to the identification of tagging behaviour, and understand their association with tagging patterns and motivations to do so.

Clustering analysis is a method that is able to split a dataset based on similarity. Typically, clustering is a non trivial task since it does not have a training set in which we could compare cluster results with predefined classes. Clustering methods for persona development rely on the use of manual identification (based on human-judgment) or semi-automated tools for data mining [BWB12]. In addition to that, depending on the issue or task goals, it can rely on the use of qualitative and/or quantitative data for fitting results. We rely on the use of clustering because of

the type of data our framework computed: besides its output resulting in quantitative data, it also has no labeling data to indicate a clear set of classes that could help us classify our data. However, the framework can provide knowledge on each user's features based on our model, which can be helpful for the identification task of tagging behavior for representative amounts of data.

6.2.1 Clustering Tools

Although clustering has been widely used for pattern recognition in data analysis, the choice of clustering algorithm is not an easy task and it does not provide a definitive answer. It depends on the type of data used and it can be decided during the process of clustering analysis. While analyzing the algorithms we could use, we first looked for those that could deal with the type of data resulted from our framework. Since we have quantitative features that are represented as continuous variables (0-1), we looked for clustering tools that could deal with this type of data. First, we looked at techniques that had been used by other works [BWB12, MLFA11] to create behavioural personas. According to the results of [BWB12], Principal Components Analysis (PCA) [AW10] combined with hierarchical agglomerate clustering showed the best results for their scenario. We took in consideration, in the decision stage, which approach to use for clustering our data. However, because we are dealing with a significant higher amount of data compared to previous works that fitted clusters for the task of modeling personas [BWB12, MSK08], we decided to resort to visualization tools to compare clustering outcomes and performance. As we created the available features to the task of tagging behaviour identification, we split the initial features (structure, language) into sub features in order to identify differences that clustering tools could reveal. Due to that, the number of features available in the framework increased, which led us to resorted to techniques that could improve the visualization of multi-dimensional data to support our decision on which clustering approach to use.

One of the most popular techniques to support the visualization task of multi-dimensional data is called Principal Components Analysis. PCA is an unsupervised learning method that consists of a data reduction technique to identify main components that support the identification of data patterns. This is a widely used technique for dimensionality reduction and it supports the explanation of principal components and how features are correlated, as well as their importance to generate relevant clusters. However, a recent approach has shown best results on the task of dimensionality reduction to support visualization. It is called t-Distributed Stochastic Neighbor Embedding (t-SNE) [MH08], and it has been used to improve visualization of high dimensional datasets. One of its advantage is that it can combine the well-known PCA technique as an additional resource to help in the task of dimensionality reduction. Such combination of techniques is suggested by the author of t-SNE as a step to help decrease data for 2D before even reducing it using t-SNE.

We relied on three aspects to support our decision on which cluster method to use: previous works that use clustering for the identification of personas [BWB12, MLFA11], comparison among different cluster methods, and availability of resources to support the identification of differences in clustering results. The first researched algorithm was K-means, that is by far the most popular tool for clustering analysis [Seg07]. K-means is a classic machine learning algorithm widely used to

support the identification of clusters in many distinct data types and structures. However, one of its drawbacks is that we need to decide on the number of clusters it would fit even if we do not know for sure how many representative clusters our data have. As we are dealing with data collected from real interactions, we decided to look for clustering methods that could allow us to skip this step and deal with data at first, without any intervention. Moreover, we know from previous research that users' tagging behaviour sometimes presents more than one motivation for tagging [GLYH10]. For this reason, we looked for a clustering algorithm that could estimate the probability of a user being part of a cluster or not, and the values referring to the features we have created to provide the possibility of interpretation of clusters vs. features. When looking for these characteristics, we found the class of Expectation Maximization (EM) algorithms, that consists of approaches to unsupervised, semi-supervised or supervised learning [FR06]. We fit a Gaussian mixture model EM algorithm, which assumes that each cluster behaves as a Gaussian, and it estimates a probability for each element to belong to a cluster.

To support our decision on which clustering algorithm to adopt, we decided to visualize how our data would change according to different clustering approaches, using t-SNE as tool for visualization. Figure 6.4 shows the outcomes of the combination of t-SNE and each one of the clustering algorithms – K-means and EM.

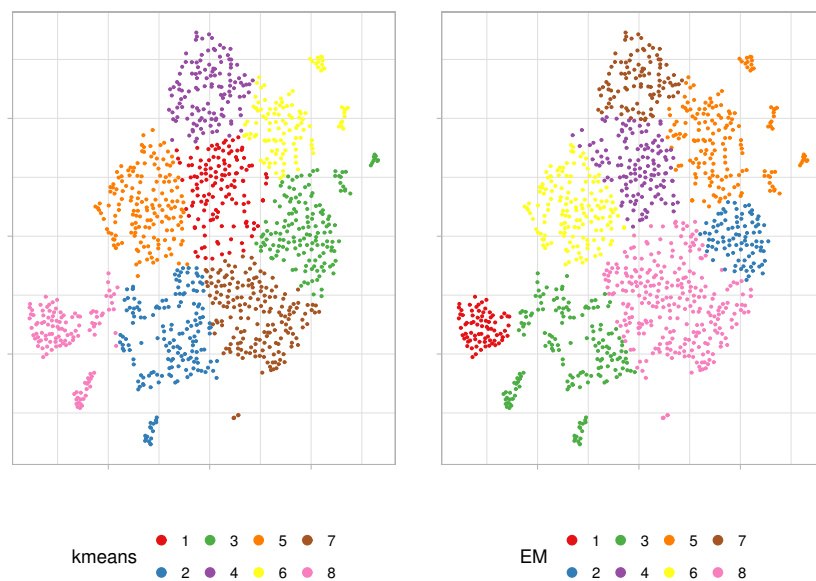


Figure 6.4: Comparison between two approaches for clustering data: K-means vs. EM.

We decided to fit a different number of clusters because K-means needs an initial parameter setting the number of clusters to fit beforehand. On the other hand, EM does not need any parameter for initialization. When running EM for the first time without fitting any number of clusters, the number of resulting clusters was nine. However, when comparing the features, which shall be discussed in detail in the next section, the clusters presented many similarities, something which is not very helpful for the creation of Personas that must be different in order to accomplish

their communication goals. As we started to decrease the number of clusters as parameters for each of the clustering approaches, what we saw was that they tend to generate similar results, with the difference that K-means presents more refined shapes, while EM presents clusters that look more like a flow.

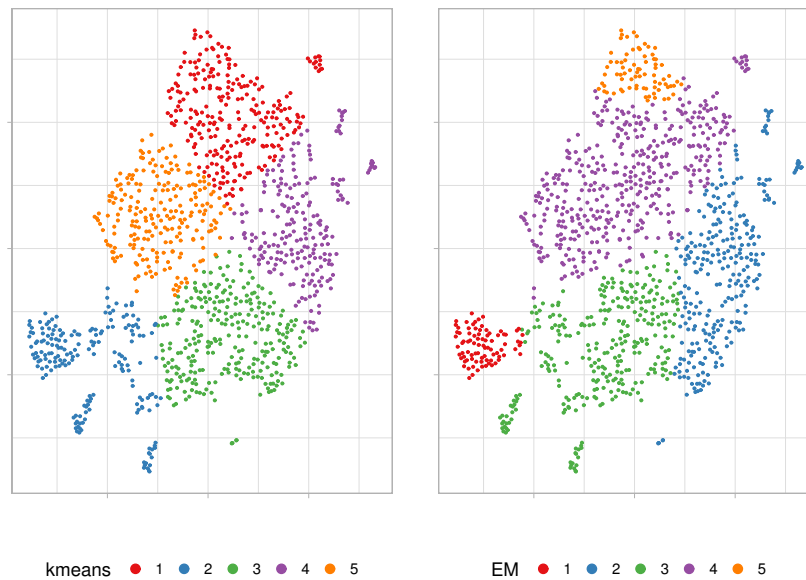


Figure 6.5: Comparing two approaches for clustering data. K-means, vs. EM.

Thus, we relied on the third aspect of clustering differences to support our decision. Through the EM algorithm, it is possible to compute probabilities for each observation to verify whether it belongs to a cluster or not. It gives us the possibility to observe data and their uncertainty, and the mean to each feature according to the cluster it belongs. This gives us a straightforward way to interpret clusters and their outcomes when looking for meaningful information originated from datasets with many features, and, for this reason, we decided to work with the EM algorithm for the clustering task. This stage of our work was only used to reach a final decision regarding what clustering approach we should use. Next, we will present the approach we use to decide among the final number of clusters to each dataset.

6.2.2 Multiple Comparisons of Features

One of the approaches used to decide on the number of relevant clusters to each dataset was a multiple comparison of features. It helps to address differences among the mean resulted from each feature of each cluster. This step is important because we wanted to have clusters that express the differences among groups and finally design meaningful personas to express the differences among tagging behaviour. At the same time, we did not want the resulting clusters to represent groups with too many similarities. To overcome this issue, we decided to conduct a Post-hoc analysis to each feature across the resulted clusters by the EM cluster algorithm. This analysis can be performed through multiple comparisons of variables and, in order to do so, we conducted a One-way analysis

of variance (ANOVA) to find the difference in the mean of features for all resulting clusters. Our null hypothesis for identifying how many clusters would be ideal for us assumes that there is no difference in the mean among the features for each cluster.

At first we started the clustering tasks without any parameters for the number of clusters, which then resulted in nine clusters. Later, we applied ANOVA and used Tukey's HSD (Honestly Significant Difference) procedure to visually compare each pair of mean by feature and then used the resulted *p-value* to analyze the differences among the features. When the difference is low between two clusters, the mean level approximates to zero and the *p-value* is higher than 0.05, as shown in the example on Table 6.4.

Figure 6.6 shows the same result from this analysis for nine clusters comparing pair by pair for the *ENG_para* feature (paradigmatic tags assigned in English). We can observe many pairs of clusters with mean levels close to zero. We so applied the same approach to other features to identify if this behaviour would also replicate. Once we identified that there are pairs of clusters with no differences, or low in mean levels, this means that the differences among the features are not very representative for the purpose of creating personas. This step supported our decision to decrease or increase the number of clusters according to the differences found among the features of each cluster. For example, since the results of nine clusters presented very low difference between the pair of features, we could step forward and decide to execute the algorithm again with fewer number of clusters as a solution to increase the differences among the clusters.

In comparison to the nine clusters we previously presented, we can observe in Figure 6.7 the difference of mean for the same feature when only five clusters were initialized by the clustering algorithm.

Also, the *p-values* presented in Table 6.5 show that only one cluster presented a significant similarity for this feature. We proceeded to identify other features that would present the same behaviour for the same pair, or other pairs that present similar behaviour. If these similarities do not repeat among the other features, or if the *p-values* show at least a relevant dissimilarity among them, we then proceed to stop the analysis of the number of clusters and start conducting the analysis of clustering results.

This approach was used whenever we were not sure about the differences among clusters. Next, we will present the resulting clusters by the EM algorithm to each of the datasets we gathered.

6.3 Findings

In this section we present the cluster results after the previous stages conducted for defining the number of clusters and the clustering algorithm we would use. As previously mentioned, we opted for the EM algorithm as a tool to support our analysis of tags from Flickr and Instagram (the latter was split into two datasets by location regarding language references). We will discuss the results found to each dataset, and how the outcomes of each cluster supported the identification of tagging behaviour and the differences among tagging motivations.

Table 6.4: *P-value* for a pair of clusters. Comparison regarding the target feature we analyzed: *paradigmatic tags assigned in English (ENG_para)* for nine clusters.

Clusters	
2-1	0.00
3-1	0.00
4-1	0.00
5-1	0.00
6-1	0.00
7-1	0.00
8-1	0.74
9-1	0.00
3-2	0.00
4-2	0.00
5-2	0.00
6-2	0.00
7-2	0.00
8-2	0.00
9-2	0.00
4-3	0.92
5-3	0.00
6-3	0.00
7-3	0.00
8-3	0.00
9-3	0.01
5-4	0.00
6-4	0.02
7-4	0.94
8-4	0.00
9-4	0.49
6-5	0.93
7-5	0.00
8-5	0.00
9-5	0.89
7-6	0.01
8-6	0.00
9-6	0.99
8-7	0.00
9-7	0.82
9-8	0.00

6.3.1 Instagram Clusters

We started our first analysis using the dataset of tags assigned mainly in English, according to the criteria of data gathering we previously described. For this analysis we kept the features that are related only to English language as sources for creating clusters. It means that, from the initial

95% family-wise confidence level

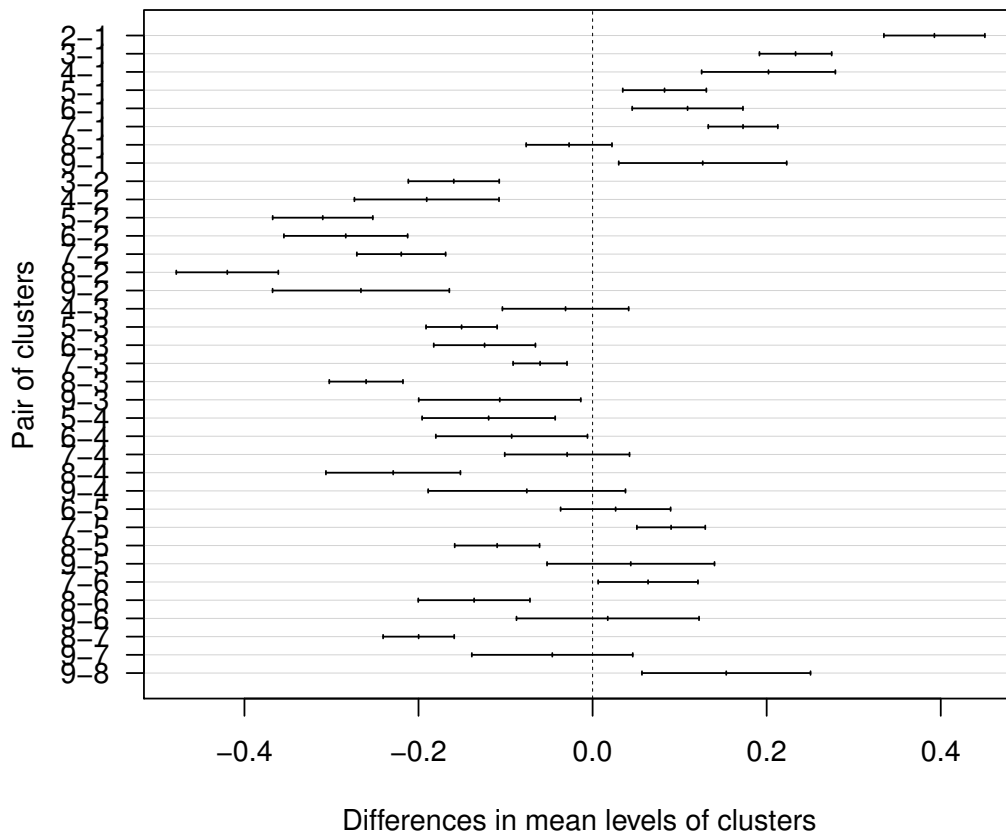


Figure 6.6: Differences in mean levels of feature *paradigmatic tags assigned in English (ENG_para)*. This feature is compared pair by pair for each of the nine fitted clusters.

Table 6.5: Multi comparison of features for five clusters. The *p-values* show that there is difference among the means resulting for the target feature being analyzed.

Clusters	p-value
2-1	0.00
3-1	0.06
4-1	0.00
5-1	0.00
3-2	0.00
4-2	0.00
5-2	0.00
4-3	0.92
5-3	0.00
5-4	0.00

number of features we had, 12 were kept at this moment. For the first iteration of the cluster algorithm, we did not fit the number of clusters it would generate, since this practice is allowed by

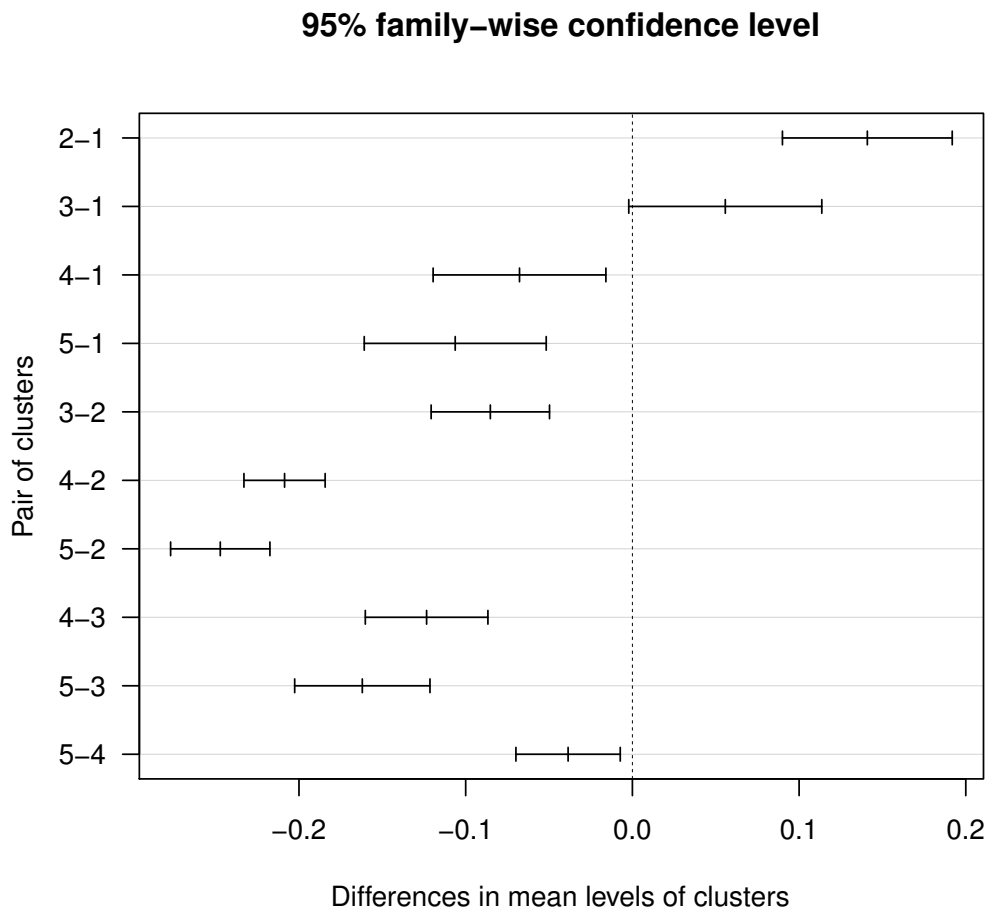


Figure 6.7: Same approach applied to evaluate the clustering performance of a different number of clusters. Fitting five clusters, the results show that the mean is not the same when comparing clusters pair by pair.

the EM algorithm. Because of that, nine clusters were resulted and we started the first analysis by watching the differences among the resulting features to each cluster, using the multi comparison of features as we have previously mentioned in Section 6.2.2. We applied Tukey's procedure and at least three clusters presented low differences in the p -value to each pair of cluster regarding its features. Due to that, we decreased the number of clusters until the cluster comparisons showed significant differences among each feature for each cluster. When the results reached five clusters, we were able to observe the differences among each result, and found representative values that present differences of tagging behaviour.

Table 6.6 shows the final clusters resulted from Instagram data, collected from regions where English is the main language spoken. The mean of each feature according to the cluster representation resulted from the clustering task shows differences of tagging behaviour. These results guided us in the identification of tagging behaviour in cases like the results for cluster #1, which we can see that the use of syntagmatic tags is transitory – very low repetition for syntagmatic tags when used, and low repetition of tags in general. Through the mean analysis of each feature and the model

we previously created, the first interpretation we assume is that users who are part of this cluster adopt the syntagmatic-contextualization dimension for tagging, which is related to self-expression and context communication. Moreover, looking at the features from cluster #2, there are a representative value of tags assigned using the paradigmatic structure more often (0.46) compared to other clusters. This could indicate that this is a user who prefers to describe the content of the image at the same time he/she uses this type of tag to index content for public audience, since rates for repetition of tags (0.59) are high and heterogeneity (0.47) is one of the lowest, among other clusters.

Table 6.6: Five clusters fitted by EM based on a Gaussian Mixture Model for the dataset of English speakers.

Variables	Cluster				
	#1	#2	#3	#4	#5
ENG_para	0.29	0.46	0.36	0.21	0.16
ENG_rep_para	0.37	0.41	0.32	0.15	0.06
ENG_het_rep_pa	0.41	0.46	0.47	0.57	0.36
ENG_Syn	0.42	0.40	0.49	0.63	0.46
ENG_rep_Syn	0.03	0.33	0.51	0.63	0.31
ENG_het_rep_syn	0.14	0.54	0.61	0.38	0.66
ENG_prop	0.60	0.74	0.74	0.76	0.53
Repeated	0.19	0.59	0.40	0.62	0.42
Rep_Var	0.71	0.47	0.60	0.41	0.56
Unknown	0.40	0.28	0.26	0.27	0.53
Unknown_rep	0.42	0.24	0.14	0.20	0.61
Unknown_het_rep	0.57	0.48	0.38	0.51	0.48

On the other hand, we see users from cluster #3 and #4 representing those users who use both structures of tags, but with higher level of repetition for syntagmatic tags. Comparing repetition in general, cluster #4 shows much more engagement in the repetition task. So we could assume that this cluster has users who want to categorize content using syntagmatic tags. Cluster #5, however, was the one that used less paradigmatic tags in English, the higher level of tags not recognized by the framework, and some level of repetition. We could assume that this cluster gathers users that use a language other than English.

Although we can assume tagging behaviour just by looking at the features, we decided to take a step forward to validate our first impressions. These results will support the stage of modeling users as Personas, as an attempt to model users' motivation for tagging and its relation with tagging behaviour in Section 6.4.

We also executed the same steps we presented before to the Instagram dataset collected from regions where Portuguese is more likely to be the main spoken language. We present the results in Table 6.7 and, differently from the previous clustering task, at this stage we analyzed tags assigned in English and Portuguese. This was a decision made based on the results of the studies we conducted with users showing that Portuguese speakers are used to assigning tags in English as well. The

results presented in the beginning of this Chapter show that this behaviour also replicates on the general numbers we computed for the Instagram dataset from Portuguese speakers' regions.

Table 6.7: Five clusters fitted by EM based on a Gaussian Mixture Model for the dataset of Portuguese speakers.

Variables	Cluster				
	1	2	3	4	5
PT_para	0.24	0.07	0.15	0.10	0.26
PT_rep_para	0.20	0.00	0.12	0.07	0.23
PT_het_rep_para	0.45	0.00	0.24	0.65	0.65
PT_synt	0.42	0.20	0.25	0.15	0.37
PT_rep_synt	0.44	0.10	0.18	0.09	0.27
PT_het_rep_synt	0.61	0.34	0.27	0.56	0.57
PT_prop	0.61	0.25	0.38	0.24	0.59
ENG_para	0.04	0.16	0.16	0.27	0.06
ENG_rep_para	0.02	0.17	0.15	0.27	0.07
ENG_het_rep_para	0.08	0.40	0.25	0.46	0.63
ENG_synt	0.08	0.27	0.19	0.26	0.09
ENG_rep_synt	0.02	0.29	0.15	0.22	0.07
ENG_het_synt	0.09	0.47	0.28	0.50	0.65
ENG_prop	0.12	0.39	0.32	0.48	0.15
Repeated	0.20	0.35	0.85	0.62	0.41
Repeated_het	0.65	0.67	0.22	0.46	0.63
Unknow	0.44	0.49	0.44	0.39	0.44
Unknow_rep	0.16	0.42	0.38	0.33	0.33
Unknow_het_rep	0.30	0.60	0.23	0.47	0.68

6.3.2 Flickr Clusters

We conducted this step of clustering analysis with the intent of understanding which differences our framework combined with clustering tools could present across different tagging datasets.

Table 6.8 shows the final clusters resulted by using the features we computed for the Flickr dataset. The results are very different from the previous dataset and give us room for discussing Flickr's users general tagging behaviour compared to Instagram users.

Differently from Instagram dataset, what we found in Flickr's clustering analysis were clusters with similar tagging behaviour that differ mainly by the language used. We believe that the reason for that is due to the way its tagging system was designed during the time the data was collected, and its target users. Flickr's target users are those intending to search or share photography among photo lovers [Sol16], or photographers. The community is less socially engaged than Instagram, and more concerned about the quality and content of images. Flickr also allows the upload of multiple images, a collection of images stored as an album, which affects the way people assign tags. These differences are identified by the analysis of the clustering results we present in Table 6.8.

By the analysis of the mean of each feature and the model of tagging behaviour we previously presented, the first interpretation we assume is that Flickr users are content-oriented, aiming at indexing content. Therefore, differently from Instagram, that shows the use of syntagmatic tags as a main tool for image contextualization, Flickr's results for this dataset show that users do not use syntagmatic structure motivated by self-expression. For example, cluster #1 presents those users that assigned tags in Portuguese and English. This cluster also brings users that assigned tags in Spanish, due to the similarity among some words that are also present in both Portuguese and Spanish dictionaries, such as amor, amigos, moto³. The resulting clusters show that some users use syntagmatic tags, but, because their general tag repetition scores high, we can assume that syntagmatic tags are not assigned with the intent of contextualizing an implicit content regarding the image subject. As we will explain later through the Persona we created, users assign many tags related to the photo location (#SanFrancisco #LatinoAmerica #NewYork), and repeat tags due to the system design that allows users to assign tags once and replicate tags for all images that are part of an album. So, it is important to verify how repetition can impact the results of clustering tools and conduct manual analysis to understand the differences among users that are close to the center of the cluster.

Despite the way Flickr's interface design affects user tagging behaviour, what we have noticed by the analysis of clustering results is that the use of paradigmatic tags is predominant among clusters and users will use tags in more than one language for the purpose of indexing content. This is another reason to reassure that Flickr is an environment where users assign tags for the task of indexing content instead of contextualization. Even when they use syntagmatic tags, e.g. cluster #2, what we found by looking at the 10 users closest to the center of the cluster was that many of them use this type of tag structure to give more information about a place, more details about a content, or to assign location names. Evidence of that is due to the low occurrence of unknown tags. Cluster #3, on the other hand, represents those users that are engaged in the task of content description. This cluster showed the lowest value for repetition of tags, and it can indicate that these users are engaged in the task of describing image and its details by a variety of different tags. Finally, cluster number #4 is the cluster that presents users who use a combination of English and other languages, such as, French, German, Spanish, among others. As we analyzed user profiles, many of the tags used were assigned in other languages and, because of that, this was the cluster with the highest mean for tags in the unknown category.

Next, we present the Personas we designed based on the cluster results we found, as well as the discussion about clusters findings and users' motivation for tagging.

6.4 Tagging-Based Personas

The combination of clustering data and user modeling has been used for supporting designers to better understand the interactions between users and systems. In the work of [BWB12], they

³amor – love; amigos – friends; moto – motorcycle

Table 6.8: Flickr clustering results for both Portuguese and English language.

Features	Cluster			
	#1	#2	#3	4
PT_para	0.24	0.00	0.00	0.00
PT_rep_para	0.24	0.00	0.00	0.00
PT_het_rep_para	0.11	0.00	0.00	0.00
PT_synt	0.14	0.00	0.00	0.00
PT_rep_synt	0.14	0.00	0.00	0.00
PT_het_rep_synt	0.08	0.00	0.00	0.00
PT_prop	0.31	0.00	0.00	0.00
ENG_para	0.29	0.69	0.92	0.58
ENG_rep_para	0.29	0.69	1.00	0.57
ENG_het_rep_para	0.09	0.11	0.16	0.14
ENG_synt	0.17	0.22	0.00	0.04
ENG_rep_synt	0.17	0.22	0.00	0.14
ENG_het_rep_synt	0.09	0.08	0.00	0.15
ENG_prop	0.47	0.92	0.93	0.62
Repeated	0.95	0.94	0.76	0.94
Repeated_het	0.21	0.13	0.18	0.11
Unknow	0.21	0.07	0.06	0.37
Unknow_rep	0.12	0.13	0.19	0.14
Unknow_het_rep	0.13	0.12	0.00	0.10

compared semi-automated clustering methods for performing user modeling through Personas. Personas are tools used for representing a set of users, their needs, goals, expertise and so on, aiming to segment groups that have similar behaviour with a task in mind. The goal of using personas is to identify general users' experiences as source for designing or redesigning products [Nie12]. In the work of Brickey et. al [BWB12], they compared clustering methods for modeling personas to experts' gold standards to evaluate their quality. The data used for clustering was gathered through a survey in which 18 participants took part. As a result, they found that quantitative data clustered PCA followed by hierarchical clustering show more accurate results in the identification of personas against qualitative data using Latent Semantic Analysis. Following this path, we expect the clustering tools we used in this case study to support us in the identification of personas as a tool for presenting the outcomes of our approach for the identification of tagging behaviour.

One of the disadvantages of using qualitative data to identify personas is that the manual identification of patterns of behaviour is challenging [BWB12]. On the other hand, quantitative persona clustering rely on algorithms and quantitative data that need a step back in order to process information in the way that quantitative information makes sense for the clustering interpretation and use.

The goal of using personas is to design a set of few archetypes, based on user research, that represent groups of users for a given scenario. The advantage of using personas to talk about differences among users is that it works as a communication tool among teams involved in a project.

Instead of resorting to numbers every time user research results are needed, personas intend to provide the general insight to support the design communication process. It is important to highlight that for the step of creating personas, the step of clustering analysis was essential in terms of grasping the general understanding of tagging behaviour. However, the manual analysis of tagging from those users closest to the clustering centers guided us to the substantial profile creating of each persona for groups of users.

Goal-Directed Perspective [Nie12] is an approach that focuses on users' work/goals to build the personas' description. In general, this approach is used with one persona in mind to design a product or service. However, because our intention is to describe users that have already been using tagging systems (as a service), we were able to identify differences in behaviour across many users through the clusters they belong to. By using this approach, we are able to explain users' attitudes and discuss outcomes and approaches based on users' research results and their intentions when using tagging systems.

Based on the research we conducted, the approach we created, and the case study results, we consider the users' communication goal as our main focus for the process of modeling personas. In order to create the personas, we resorted to the elements that were closer to the center of the cluster. We tried to accommodate, whenever possible, the similarities of users from Portuguese and English speaking countries but, due to differences among languages in some cases, such attempt was not possible. As final results we present a compilation of personas that could be identified through the analysis of clustering results in combination with the manual analysis of user features resulting from the framework application and the set of tags to each one of the users closest to the clusters' center.

6.4.1 Personas for Instagram dataset

Our findings show that although clusters in both datasets have similarities, a manual analysis to understand the differences among the results was necessary because of the features' means resulted from each dataset presented differences. Clusters and features are guides for our final conclusion, but looking at the data, repeated tags and user profiles supported us in the stage of modeling personas and their characteristics. As we consider each clustering result, we decide to access the 10 closest elements from the center of each cluster to support us in the stage of modeling users' tagging behaviour. The data analysis was based on individuals' features, the set of repeated tags and the visualization of user profiles on Instagram. Table 6.10, 6.9 and 6.11 present the Personas we found based on the analysis of the clustering results from Portuguese and English speakers' dataset. Because we found similarities among the clusters, we accommodated together those personas from different datasets that presented similarities in their tagging behaviour. As their behaviour was similar even having different means, we decided to discuss their differences as we present them. Next, we discuss the characteristics of each created persona to each one of the clusters.

The Social: a persona that represents those users that use tags for self-expression. As we analyzed users close to the center, the use of repetition of syntagmatic tags was almost exclusively

to refer to events, like parties. Its occurrence was very low, and for users close to the center, many times it was null. Repetition, as we can observe in the cluster chart, represents the lowest mean compared to other clusters. It means that this user is not interested in categorization of content. Tags are transitory and used for the moment. For the Portuguese dataset, the mostly used language is Portuguese, which can indicate that it is used basically to communicate with the followers the users know. Clusters: #1 – Portuguese speakers; #1 – English speakers.

The Social Categorizer: as the name says, this persona represents those users that use tags aiming at categorizing context. They use tags in English and Portuguese, and syntagmatic tags generally prevail as preferred. There is low use of repetition, which occurs according to image context. It scores high for the use of tags for later browsing (#friendsfromcollege) that brings reference of the subject. It eventually uses paradigmatic tags, mostly in English. Repetition of tags in English could indicate a restrict vocabulary. We call them social because of the use of syntagmatic tags, low repetition and high heterogeneity of tags generally speaking. Clusters: #2 – Portuguese speakers;

The Entrepreneur: – this persona represents those users that have high repetition of tags due to the central goal of sharing content. To our surprise, the cluster brought together mainly users with the intent of categorizing content and promoting brand awareness, reaching consumers or promoting consumers' engagement. Regarding the Portuguese speakers' dataset, the use of tags in English and Portuguese is well balanced, which shows that these users are trying to expand their audience by using tags in another language. This was the persona that presented the highest value for repeating tags in general, with the lowest value for heterogeneity. Clusters: #3 – Portuguese speakers; #4 – English speakers.

The Expert: this persona represents those users that could be called tagging experts. They know tagging functions and use a variety of tagging structures according to what they want to express or the audience they want to reach. They could be considered a mix of the describer, the categorizer, and the social personas. Besides that, they use tags in English and Portuguese. We can observe in the clustering results that the users representing this cluster will score high for repetition of syntagmatic tags, and present one of the highest values to this feature. Also, heterogeneity for syntagmatic repeated tags is high. This could indicate that at the same time one has been using tags for categorization, the same structure is being used for contextualization and represents a well defined vocabulary with this purpose. Clusters: #4 – Portuguese speakers;

The Social Describer: as the name says, this persona represents those users that use tags with the intent of describing the content of an image with indexing purposes. Because content is more important than context, these users are more likely to represent those users with high scores for repetition and paradigmatic tags. However, for our surprise, when analyzing cluster results and users close to the center, we also found syntagmatic tags assigned to their images. We call them social describers due to the fact that at the same time this persona is concerned with content indexing, he/she also uses syntagmatic tags to express the image context, when appropriated. This persona differs from the categorizer persona (#3 – English speakers) mainly by the predominant occurrence

of paradigmatic tags, while categorizers use more syntagmatic tags with higher repetition. Clusters: #2 – English speakers; #5 – Portuguese speakers.

The Categorizer: This persona differs from the previous one basically because its behaviour is more related to the categorization of content by the use of syntagmatic tags. Repetition of syntagmatic tags is high, and its heterogeneity is also high. He/she also uses paradigmatic tags but with less frequency than the previous persona. He/she does not use any other language than English, but the vocabulary is clean and well recognized by the framework. He/she differs from the “Social Categorizer” because of the high use of paradigmatic tags and indication that indexing is more important than contextualization. Clusters: #3 – English speakers.

The Teenager: this persona represents users that know the language of the internet. They could be compared to people classified as part of the Generation Z. These users are always aware of the latest trend on the internet, new memes, acronyms to express context about a topic, image or content. They use it as a fast way to self-expression. Many unknown tags were found in their cluster because these are expressions that are created everyday, many of them only available in the internet language, such as, #TGIF (thank God is Friday), and #OOTD (outfit of the day), #instadog (referring to photos of dogs posted on Instagram). Clusters: #5 – English speakers.

6.4.2 Personas for Flickr dataset

For the task of creating Personas for Flickr, we followed the same steps described for the Personas created for Instagram dataset. Our findings show similarities among the clusters fitted to Flickr dataset, but we tried to address the differences in the Personas presented in Table 6.13. Differently from the previous Instagram dataset we have analyzed, we do not have any information regarding users’ location or language of preference according to the region where data were gathered. The results found regarding the adopted language were purely based on the libraries we used, the model we have based our framework on, and the analysis of tagging behaviour presented by the clusters fitted during the case study we have conducted. One important point to highlight is that the way Flickr interface is designed could be the reason for similar tagging behaviour among clusters. What we have found is that all clusters have indexing of content as main motivation for tagging. This is reflected in the Personas’ goals we pointed out, and even such goals being the same, we still could detect differences on tagging behaviour by the identification of differences among the language adopted for clustering in each cluster. Next, we present these differences across the Personas’ description we have created.

The Traveler – Multilanguage: This persona represents those users that use a combination of Portuguese and English to assign tags to images. Besides that, syntagmatic and paradigmatic tags were also presenting in both languages. Syntagmatic tags were used to assign names of places, and we can observe that due to their repetition and low heterogeneity. Among users that represent this persona, we also found users that assign tags in Spanish. The reason for that is that some tags in Spanish are similar to words in Portuguese. This was one of the limitations we found due to dataset lack of information about users’ geolocation. Cluster: #1.

The Traveler – English: This persona represents users who are English speakers and use tags with the purpose of describing image content (paradigmatic), and give more details about location (syntagmatic), both with indexing purposes. High repetition of syntagmatic tags indicates that it may be affected by the interface design. In other words, the occurrence of syntagmatic tags here does not indicate that this persona has contextualization motivations, or even that repetition is due to categorization intentions. However, categorization of content occurs naturally as he/she gives more information about the image content through syntagmatic tags. Cluster: #2.

The Describer: This persona represents those users that use paradigmatic tags for describing image content. We can identify these users as those who use mainly paradigmatic tags. There is a variability in repeated tags, pointing to the fact that these users try to describe image content as much as they can. This also can indicate a variability in the image subject, and they may not use the album resource for photo uploading. Vocabulary is very clear, and English is the only language used for tagging. Cluster: #3.

The Foreigner: This persona represents users that use foreign language to assign tags alongside English. We found users that have assigned tags in French, German, and Spanish. The unknown feature is the indicator of that characteristic in the clustering results. One interesting point is that this persona also uses English alongside other languages, the same behaviour found for Portuguese speakers during user studies. This can indicate that other cultures may also present the same tagging language behaviour we modeled in our language approach for this work. Cluster: #4.

⁴Zwiebel – onion translated from German.

⁵Schwul – humid translated from German

Table 6.9: Personas based on clustering results from Instagram datasets.


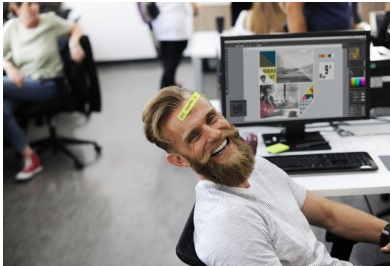

Behaviour	Personality
	<p>The Social</p> <ul style="list-style-type: none"> ▪ Goal: contextualization; ▪ Audience: personal; ▪ They use tags for social interaction. Self-expression through tags that give details about the image context (#transformationtuesday) and the subject matter. Posting event pictures is a must and they will repeat tags in these situations. Very low repetition though. Descriptive tags, when used, are for stressing the context. Tags: #goldenbirthday #best-friends. <p>Clusters: #1 Portuguese speakers; #1 English speakers;</p>
	<p>The Social Categorizer</p> <ul style="list-style-type: none"> ▪ Goal: context indexing and personal organization; ▪ Audience: personal; ▪ They use tags for personal organization aiming at categorizing images by subject. Due to that, tags will sometimes have social nature (#friendfromcollege). They will use unique tags. Content description is not their priority, although they will use it with caution and in English. They are aware that there is an audience that could be reached by the use of foreign language, so they will engage in this practice occasionally. Tags: #marina40 #juliosilva <p>Cluster: #2 Portuguese speakers.</p>
	<p>The Entrepreneur</p> <ul style="list-style-type: none"> ▪ Goal: content indexing (for Business); ▪ Audience: public; ▪ It scores high for small companies trying to reach consumers and promote brand engagement. Tags are for future retrieval, self-reference and spreading content. It presents the highest repetition of tags for keeping track of content. Tags: #slamrichmond, #dryerase-tracks #pomegranateinn, #thewowfactorcakes <p>Cluster: #3 – Portuguese speakers, #4 – English speakers.</p>

Table 6.10: Personas based on clustering results from Instagram datasets.



Behaviour	Personality
	<p>The Expert</p> <ul style="list-style-type: none"> ▪ Goal: indexing & contextualization; ▪ Audience: it depends on the tag, moment, or goal; ▪ The expert knows when, how, and why to use tags. He knows that tags are powerful tools that can be used according to his goal when sharing an image. Context indexing occurs naturally through the reuse of some tags to self-expression. <p>Cluster: #4 Portuguese speakers;</p>
	<p>The Social Describer</p> <ul style="list-style-type: none"> ▪ Goal: indexing; ▪ Audience: public; ▪ They know how to use tags for expanding visibility of images. Tags have high repetition with substantial variability. They use personal categorization. Common sense words or key-words. They are photographers and professionals trying to reach a broad audience for their content. Engagement in contextualization of content, but as a secondary behaviour. Contextualization of content can be related to the location where the images were taken. Tags: #dog, #sky #gopro #sunny #goodmorning #newyorkcity. <p>Clusters: #2 English speakers; #5 Portuguese speakers;</p>

Table 6.11: Personas based on clustering results from Instagram datasets.




Behaviour	Personality
	<p data-bbox="627 640 836 674">The Categorizer</p> <ul style="list-style-type: none"> <li data-bbox="671 703 903 736">▪ Goal: indexing; <li data-bbox="671 770 1254 804">▪ Audience: personal with public awareness; <li data-bbox="671 837 1390 1133">▪ They use tags with the purpose of content organization. Tags bring contextualization, but the use of paradigmatic tags indicates their indexing purpose is more intense than the “Social Categorizer”. They could be considered as experts as well, because of their well-defined vocabulary and the combination of description of images alongside categorization and contextualization. <p data-bbox="703 1155 1091 1189">Cluster: #3 English speakers.</p>
	<p data-bbox="627 1234 922 1267">The Teenager (Gen Z)</p> <ul style="list-style-type: none"> <li data-bbox="671 1301 1007 1335">▪ Goal: context indexing; <li data-bbox="671 1368 1198 1402">▪ Audience: public (likes) and personal; <li data-bbox="671 1435 1390 1693">▪ They know every meme, acronym or statement around social media. They use it to express the context of the image. Repetition is due to reuse of tags, such as, #nofilter, #TGIF, #TBT, considered as unknown. Index content by putting it in a pool of images that has the same related context. Scores high for getting likes due to tags. <p data-bbox="703 1715 1091 1749">Cluster: #5 English speakers.</p>

Table 6.12: Personas for Flickr dataset we have analyzed.

Behaviour	Personality
	<p>The Traveler – Multilanguage</p> <ul style="list-style-type: none"> ▪ Goal: indexing; ▪ Audience: public; ▪ Language: English, Portuguese, Spanish; ▪ They go to places, use more than one language for indexing content across different audiences looking for the same subject or content. A mix of syntagmatic and paradigmatic tags for content description and specification of location. Syntagmatic here is not used for self-expression. Presence of high repetition of tags. Tags: #motos #viaje #CentralAmerica #friends
	<p>The Traveler – English</p> <ul style="list-style-type: none"> ▪ Goal: indexing; ▪ Audience: public; ▪ Language: English; ▪ They are explorers and use tags for content description and specification of location. Due to high repetition, contextual tags are used for categorization of content. English is their main language for tagging and they do not use any other language for this task. Tags: #newyork #winterolympics #torino
	<p>The Describer</p> <ul style="list-style-type: none"> ▪ Goal: indexing; ▪ Audience: public; ▪ Language: English; ▪ They share images that have content as their main focus. They use tags to reach people that look for images with specific content. Sharing and finding content are their main goal.

Table 6.13: Flickr Personas

Behaviour	Personality
	<p>The Foreigner</p> <ul style="list-style-type: none"> ▪ Goal: indexing; ▪ Audience: public; ▪ Language: English, and others; ▪ They want to index content and reach different audience by the language used. English is not their first language, but they will use it for indexing purposes. They use a lot of tags in English, which shows that, even though it is not their mother language, they will still use it due to its indexing advantages. High scores for the use of acronyms and for naming locations. Tags: #zwiebel⁴ #schwul⁵

7. FINAL CONSIDERATIONS

Tagging is a popular tool for classifying content online and it has being widely adopted to support search engines on the task of indexing content, users' personal content organization, and social navigation [GLYH10]. As a result of its use, tagging systems has been investigated from the point of view of users' motivation, and the type of tags they adopt. Although many researchers have conducted manual analysis to identify users' motivations for tagging, little has been investigated regarding tagging patterns that could support the identification of tagging behaviour by quantitative features. With this goal in mind, the questions addressed in this thesis aimed at finding tagging patterns that could support the identification of tagging behaviour [ZBS16a,ZBS16b]. To answer our first research question (RQ1 – Are there any patterns of tags that can contribute to the identification of tagging behaviour?), we conducted two user studies with participants from Canada and Brazil. By conducting experiments with 91 participants (34 from Canada, and 57 from Brazil) in two different conditions (with and without recommendation support), we were able to identify that language and structure of tags are patterns that point to users' communication intentions when tagging. By comparing results from both stages, we found that the type of image being tagged affects the way users assign tags. We modeled the results we found using a semiotic approach that classifies tags based on their structure – paradigmatic or syntagmatic. Images classified with high context involved are prone to receive more syntagmatic tags to express more than the content they have. When tagging these types of images, users are motivated by image contextualization. It means that content indexing is not the user's priority in this case. Instead, the message the tag expresses will refer to what is happening in the image and not what the image has.

On the other hand, when users are motivated by indexing image content, they will use paradigmatic tags in order to describe the image with units of words that represent its content, instead of its context/subject. Moreover, regarding language choice for tagging, users will switch from one language to another according to their goals and audience. This was a behaviour we found mainly for participants from Brazil, while participants from Canada presented a totally different behaviour – they assigned tags in another language mainly when suggested by the recommendation system and when image location was explicitly presented.

In order to deeply investigate these differences and users' motivations for tagging (RQ2 – How are patterns of tags, regarding structure and language, related to users' motivation for tagging?), we resorted to open-ended questions to collect users' point of view regarding their reasons for choosing syntagmatic or paradigmatic structure, and their language preference for tagging [ZMS]. As we started modeling tagging patterns against the motivations we found in the qualitative study, and based on the literature review we conducted, we concluded that users' choices for tagging are related to two main motivations – contextualization and indexing. We decided to create a framework to quantify tagging patterns and further support the identification of tagging behaviour related to users' motivation with the support of clustering tools and two datasets of tags.

Understanding how users perform the same task in different environments can provide insight for designers to decide among distinct approaches according to users and system needs. The case study we conducted in this work can be considered as a substantial behaviour from real tagging datasets and it was a necessary stage to help us answer RQ3 (Is it possible to automatically identify tagging patterns to support the identification of tagging behaviour?). Such case allowed us to identify common tagging behaviour that, otherwise, would not be feasibly possible to be manually performed by designers while analyzing such amount of data. We were able to point to the differences among tagging behaviour and how the choice of structure or language for tagging could be used as source to identify users' motivation for tagging when sharing content online. By using clustering tools we found that the patterns and motivation we have modeled replicate in the clusters and their features' behaviour. We found that the use of syntagmatic tags are the preferred structure for those who want to contextualize and categorize an image. Categorization was one of the behaviours we have considered inside the indexing motivations of our model, but we were not able to identify in the experiments we conducted due to lack of repetition of tags. By conducting the case study with real world datasets, these differences emerged. The use of syntagmatic tags for contextualization and categorization differs by their repetition and heterogeneity. We also found evidence in the Flickr dataset that other cultures that are not composed by English native speakers may have the same tagging behaviour we found for Portuguese speakers regarding the use of tags in a foreign language – English in combination with their mother language. These outcomes were considered as quantitative evidence that it is possible to identify tagging behaviour and users' motivation by computing tagging patterns as features. However, an important point to mention is that the combination of personal insight and the outcomes from the framework were essential in the task of identifying users' specific goals when repetition is involved. For example, one unexpected finding revealed that when a user assigns tags with the goal of indexing content, he/she may be promoting a brand, a blog, or a product. Although we noticed that, in this case, this was the cluster with the highest mean for repetition, we were only able to identify that this cluster was motivated by promoting brand engagement when we look at the set of tags repeated for users closer to the center of the cluster.

These were some of the behaviours we found in our work during the analysis of data gathered from Instagram that allowed us to evidence some aspects of tags, such as:

- The structure of tags in combination with measures regarding repetition and heterogeneity are essential to better understand tagging behaviour.
- Users use more than one structure for tagging. For example, users who are aware of different outcomes their tagging choice can produce will present a well-balanced use of different structures, which will be reflected in the set of repeated tags.
- Instagram has an environment that promotes users' engagement through the use of tags. With this evidence in mind we were able to differ those profiles that use tags to contextualize images

with the intent of engaging into self-expression to promote socialization, and those users who are using tags to promote content by indexing their brand and image content.

- Users assign paradigmatic tags by nature, although we did find some users that do not assign any paradigmatic tags – for example, closest users to the center of cluster 1, The Socials – there were still some users in this cluster who assign this type of tag.

We also were able to identify differences among datasets and how these reflected in users' patterns of tags and behaviour. Flickr, differently from Instagram, is generally used by users who want to promote their images because of the content they have, not their context. This goal reflects in the way the system was modeled but, based on the results we found, it may be failing to promote diversity on the use of tags to promote content. We assumed that because of the way users assign tags – supported by a tool that allows assigning a set of tags that replicates in a collection of images – the heterogeneity of repeated tags was low. It is important to point that even though it could result in good organization outcomes for users' images, the system may be losing valuable information of images to promote content indexing. However, the structure used for tagging combined with the language identification supported us in the identification of differences among groups of users and in the interpretation of tagging behaviour in a dataset, where all users seem to have the same goal in mind when using tags.

Regarding the language choice for tagging on Instagram, our approach addressed this behaviour mainly on tags assigned by users located in Brazil during the clustering task. They presented a tendency to assign tags in English and some clusters show this behaviour very clearly. This behaviour was also identified in Flickr's dataset. Although we did not know beforehand if languages other than English and Portuguese were present in the dataset, by observing clustering results we were able to identify a cluster represented as "The Foreigners", those who use other languages in combination with English.

Therefore, based on the results of our case study, what we noticed is that one of the main advantages of our model is that, by computing the features we presented, it was possible to identify tagging behaviour by looking at quantitative data to support insights about assigned tags. Through our approach, we provide a guide for the identification of tagging patterns and use this information as resource for computing features that support the identification of tagging behaviour. Designers may use our approach to identify tagging behaviour and decide which type of tags are more beneficial to their system goals, or to keep users into social engagement. Although we know that in general recommendation algorithms are quantitative approaches that use collective knowledge for recommending content, an analysis of tagging behaviour and stakeholders' goals for the system could be put in place to seek understanding of the consequences of recommendation and the outcomes it will raise. To this end, our approach could benefit designers when selecting data for recommendations based on tagging behaviour, processing data for tagging recommendation based on users profile, and also as a source to support user-centered design.

7.1 Limitations

As far as we know, this is the first work that approaches the tagging task from a semiotic point of view. Due to this fact, this work still needs some investigation in language identification and would have had better outcomes if more than one language could have been analyzed beside English and Portuguese. Besides that, once users create new acronyms on the internet from statements that sometimes cannot be identified in any dictionary, we could have achieved better results regarding users' motivation for tagging if tags such as "TBT"¹ had not been considered as unknown. This could be solved by creating a common dictionary of popular tags across the internet, highlighting their goals in terms of use and meaning.

Another limitation this work has faced is related to the choice of images to each stage of the experiment. Unfortunately, we did not use images that provide location information during the stage when users did not have the support of recommendation. It would be interesting to compare the outcomes of tags regarding this type of images and verify if they would have had more syntagmatic tags involved in the first stage and less location-based tags.

In addition, another point that could be improved is the way we computed the features used in the clustering task. Although proportion helped us understand users' general behaviour when tagging, it may overestimate the use of tags when the sample of tags is too small. For this reason, proportion should be used with caution and managed with a relevant number of tags that could give a general understanding of users' tendency regarding tagging patterns.

7.2 Future Works

Throughout this thesis we worked with the identification of tagging behaviour from a language perspective. Some future directions could use the combination of tagging patterns we found here to address more specific questions regarding users' behaviour in connection to other users' characteristics, such as the number of followers, likes based on tagging structures, languages, among others. In addition, we summarize a set of future works that we have in mind to refine the results we found:

- *Evaluation*: presenting the results we found to designers so they could raise questions about the personas we created and help refine their description.
- *Dimensions*: exploring other dimensions that could be useful for the task of identifying tagging patterns, such as the identification of Conceptual tags.
- *Recommendation*: exploring how the previous selection of tagging patterns would affect users' choice when syntagmatic tags are more present in recommendations than paradigmatic ones.
- *Applications*: investigating the same tagging patterns in different social media system or environments, such as Blogs, Bibsonomy, Tumblr, Twitter, Facebook, etc.

¹Throwback Thursday

- *Languages*: exploring the same dimensions in other languages to compare tagging from a broader point of view.
- *Measures*: investigating other measures to compute features and compare how they could refine the results we found. The ideal would be to find measures that best represent users' behaviour and still be able to use them to get insight on user individual tagging behaviour and also as a group with the support of clustering tools.

8. Activities During Period as Ph.D. Student

During the time spent at PUCRS and Dalhousie University as a Ph.D. student, besides the results we achieved with this work, we also developed correlated projects in the area of Human-Computer Interaction:

Patent

During the first year of the Ph.D. program, the author of the thesis participated in a project funded by HP – Hewlett Packard research lab at PUCRS. The outcomes of the project resulted in a service registered by the company that was granted in September 2017.

- M. Riss, N. Venkata, R. Chamun, J. de Oliveira, I. Manssour, A. de Carvalho Alvarez Ziesemer, Displaying a folding document, US Patent 9,772,977 (Sep. 26 2017).

Awards

First place in the IHC Design Competition – 2015, Public Visualization to Improve Cities. XIV Brazilian Symposium on Human Factors in Computer Systems.

Visiting Student

Conducted research and the second stage of the users' studies for this thesis and participated in the Hypertext Augmenting Intelligent Knowledge Use (HAIKU) research group under the supervision of Professor James Blustein at Dalhousie University – Halifax, Canada.

Publications

- A. de CA Ziesemer, J. B. S. de Oliveira, Keep querying and tag on: Collaborative folksonomy using model-based recommendation, in: International Conference on Collaboration and Technology, Springer, 2013, pp. 10-17.
- A. Ziesemer, L. Muller, and M. Silveira. Gamification aware: users perception about game elements on non-game context. Proceedings of the 12th Brazilian Symposium on Human Factors in Computing Systems. Brazilian Computer Society, 2013.
- A. Ziesemer, L. Muller, and M. Silveira. Just rate it! gamification as part of recommendation. International Conference on Human-Computer Interaction. Springer, Cham, 2014.

- C. Santos, A. Ziesemer, L. Espindola, P. Pires, L. Muller, and M. Silveira,. How Can I Help You? Preliminary Studies About User Strategies and Preferences During a Game. Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems. Brazilian Computer Society, 2015.
- A. Ziesemer, J. Blustein, M. S. Silveira, Multilingual tagging behaviour: The role of recommender systems, in: Extended Proceedings of the 24th Conference on User Modeling Adaptation and Personalization, UMAP'16, 2016.
- A. Ziesemer, L. Muller, M. Silveira, More than content classification: Self-expression through image tagging, in: 15th International Conference WWW/Internet, 2016.
- A. Ziesemer, J. Blustein, M. Silveira, Users tagging behavior and the effect of recommendation,in: Proceedings of the 15th Brazilian Symposium on Human Factors in Computer Systems, IHC'16, ACM, New York, NY, USA, 2016, pp. 36:1 – 36:4.
- A. Schunk, F. Bergmann, R. Piccoli, A. Ziesemer, I. Manssour, J. Oliveira, and M. Silveira. User Impressions About Distinct Approaches to Layout Design of Personalized Content. In Information Technology: New Generations, pp. 1009-1020. Springer, 2016.
- L, Oliveira, L. Espindola, C. Santos, A. Ziesemer, L. Muller, and M. Silveira, Help Resources in Games: Gamers' Opinions and Preliminary Design Remarks. Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems. Brazilian Computer Society, 2017.

Participation as a reviewer in the following committees

- Long-paper reviewer – CHI 2017;
- Long-paper and short-paper reviewer – IHC 2017;
- Design Competition – IHC 2017;

Bibliography

- [ACdS⁺14] Araújo, C. S.; Corrêa, L. P. D.; Silva, A. P. C; Prates, R. O.; Meira, W. “It is not just a picture: revealing some user practices in instagram”. In: 9th Latin American Web Congress, 2014, pp. 19–23.
- [AN07] Ames, M.; Naaman, M. “Why We Tag: Motivations for annotation in mobile and online media”. In: SIGCHI Conference on Human Factors in Computing Systems, 2007, pp. 971–980.
- [AT05] Adomavicius, G.; Tuzhilin, A. “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17-6, Apr 2005, pp. 734–749.
- [AW10] Abdi, H.; Williams, L. J. “Principal component analysis”. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2-4, Jul 2010, pp.433–459.
- [BOH12] Bostandjiev, S.; O’Donovan, J.; Höllerer, T. “Tasteweights: a visual interactive hybrid recommender system”. In: 6th ACM Conference on Recommender Systems, 2012, pp. 35–42.
- [BS97] Balabanović, M.; Shoham, Y. “Fab: content-based, collaborative recommendation”. *Communications ACM*, vol. 40-3, Mar 1997, pp. 66–72, .
- [BWB12] Brickey, J.; Walczak, J; Burgess, T. “Comparing semi-automated clustering methods for persona development”. *IEEE Transactions on Software Engineering*, vol. 38 -3, June 2012, pp. 537–546.
- [Cha00] Chandler, D. “Semiotics for beginners”. University of Wales, 2006, 206p.
- [Con17] Constine, J. “Instagram’s growth speeds up as it hits 700 million users”. Retrieved from: <https://techcrunch.com/2017/04/26/instagram-700-million-users/>, Apr 2017.
- [CS08] Corbin, J.; Strauss, A. “Basics of qualitative research: techniques and procedures for developing grounded theory”. Sage Publishing, 2008, 456p.
- [dCZdO13] Ziesemer, A. C. A; Oliveira, J. B. S. “Keep querying and tag on: Collaborative folksonomy using model-based recommendation”. In: International Conference on Collaboration and Technology, 2013, pp. 10–17.
- [DF10] Dong, W.; Fu, W. T. “Cultural difference in image tagging”. In: SIGCHI Conference on Human Factors in Computing Systems, 2010, pp. 981–984.

- [DFT10] Dattolo, A.; Ferrara, F.; Tasso, C. "The role of tags for recommendation: a survey". In: 3rd International Conference on Human System Interaction, 2010, pp. 548–555.
- [DS05] De Souza, C. S. "The semiotic engineering of human-computer interaction". MIT press, 2005, 318p.
- [DS11] De Saussure, F. "Course in general linguistics". Columbia University Press, 2011, 289p.
- [EG12] Eleta, I.; Golbeck, J. "A study of multilingual social tagging of art images: Cultural bridges and diversity". In: Conference on Computer Supported Cooperative Work, 2012, pp. 695–704.
- [EK10] Easley, D.; Kleinberg, J. "Networks, crowds, and markets: Reasoning about a highly connected world". Cambridge University Press, 2010, 42p.
- [Eve05] Evert, S. "The statistics of word co-occurrences". PhD thesis, Dissertation, Stuttgart University, 2004, 353p.
- [Far17] Farnsworth, M. "For 10 years, the twitter hashtag has fueled both social activism and dad jokes". Retrieved from: <https://www.recode.net/2017/8/23/16180180/ten-year-anniversary-birthday-twitter-hashtag-started>, Aug 2017.
- [FR06] Fraley, C.; Raftery, A. E. "Mclust version 3: an r package for normal mixture modeling and model-based clustering". Technical report, Washington University Seattle Dep. of Statistics, 2006, 58p.
- [GGQJ13] Gavilanes, R. G.; Quercia, D.; Jaimes, A. "Cultural dimensions in twitter: Time, individualism and power". In: 7th International AAAI Conference on Weblogs and Social Media, 2013, pp. 195–204.
- [GH06] Golder, S. A.; Huberman, B. A. "Usage patterns of collaborative tagging systems". *Journal of information science*, vol. 32, Apr 2006, pp. 198–208.
- [GKE11] Golbeck, J.; Koepfler, J.; Emmerling, B. "An experimental study of social tagging behavior and image content". *Journal of the Association for Information Science and Technology*, vol. 62, May 2011, pp. 1750–1760.
- [GLYH10] Gupta, M.; Li, R.; Yin, Z.; Han, J. "Survey on social tagging techniques". *ACM SIGKDD Explorations Newsletter*, vol. 12, Jun 2010, pp. 58–72.
- [Gri16] Griffin, A. "Facebook posts becoming less personal as site looks to encourage people to post about their lives". Retrieved from: <https://goo.gl/ckECxx>, Jul 2016.

- [HTE10] Huang, J.; Thornton, K. M.; Efthimiadis, E. N. "Conversational tagging in twitter". In: 21st ACM Conference on Hypertext and Hypermedia, 2010, pp. 173–178.
- [hun] "Hunspell". Retrieved from: <http://hunspell.github.io/>, January 2015.
- [Jen] Jenks, G. "Wordsegment". Retrieved from: <https://pypi.org/project/wordsegment/>, January 2015.
- [JSK10] Jawaheer, G.; Szomszor, M.; Kostkova, P. "Characterisation of explicit feedback in an online music recommendation service". In: 4th ACM Conference on Recommender Systems, 2010, pp. 317–320.
- [Kel] Kelly, R. "Pyenchant". Retrieved from: <https://pypi.org/project/pyenchant/>, January 2015.
- [KKG10] Körner, C.; Kern, R.; Grahsl, H. P.; Strohmaier, M. "Of Categorizers and Describers: An evaluation of quantitative measures for tagging motivation". In: 21st ACM Conference on Hypertext and Hypermedia, 2010, pp. 157–166.
- [KKL17] Kowald, D.; Kopeinik, S.; Lex, E. "The Tagrec Framework as a Toolkit for the Development of Tag-based Recommender systems". In: 25th Conference on User Modeling, Adaptation and Personalization, 2017, pp. 23–28.
- [KPL17] Kowald, D.; Pujari, S. C.; Lex, E. "Temporal Effects on Hashtag Reuse in Twitter: A Cognitive-inspired Hashtag Recommendation Approach". In: 26th International Conference on World Wide Web, 2017, pp. 1401–1410.
- [KRW11] Knijnenburg, B. P.; Reijmer, N. J. M.; Willemsen, M. C. "Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems". In: 5th ACM Conference on Recommender Systems, 2011, pp. 141–148.
- [KT03] Kelly, D.; Teevan, J. "Implicit feedback for inferring user preference: a bibliography". *ACM SIGIR Forum*, vol. 37-2, Sep 2003, pp. 18–28.
- [Lac17] Lachowicz, D. "Abiword/enchant". Retrieved from: <https://github.com/AbiWord/enchant>, August 2017.
- [Lay15] Layton, R. "Learning Data Mining with Python". Packt Publishing Ltd, 2015, 369p.
- [LB11] Lui, M.; Baldwin, T. "Cross-domain feature selection for language identification". In: International Joint Conference on Natural Language Processing, 2011, pp. 553–561.
- [LFH17] Lazar, J.; Feng, J. H.; Hochheiser, H. "Research methods in human-computer interaction". Morgan Kaufmann, 2017, 560p.

- [Max08] Maxwell, J. A. "The SAGE handbook of applied social research methods: Designing a qualitative study". SAGE Knowledge, 2008, 214p.
- [MH08] Maaten, L. V.; Hinton, G. "Visualizing data using t-sne". *Journal of Machine Learning Research*, vol. 9, Nov 2008, pp. 2579–2605.
- [MKB⁺16] Mahajan, D.; Kolathur, V.; Bansal, C.; Parthasarathy, S.; Sellamanickam, S.; Keerthi, S.; Gehrke, J. "Hashtag recommendation for enterprise applications". In: 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 893–902.
- [MLFA11] Masiero, A. A.; Leite, M. G.; Filgueiras, L. V. L.; Aquino Jr, P. T. "Multidirectional knowledge extraction process for creating behavioral personas". In: 10th Brazilian Symposium on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction, 2011, pp. 91–99.
- [MSK08] Miaskiewicz, T.; Sumner, T.; Kozar, K. A. "A latent semantic analysis methodology for the identification and creation of personas". In: SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 1501–1510.
- [Net80] Netto, J. T. C. "Semiótica, informação e comunicação: diagrama da teoria do signo". Editora Perspectiva, 1980, 222p.
- [Nie12] Nielsen, L. "Personas-user focused design". Springer Science & Business Media, 2012, 154p.
- [OLL08] Oh, J.; Lee, S.; Lee, E. "A user modeling using implicit feedback for effective recommender system". In: Conference on Convergence and Hybrid Information Technology, 2008, pp. 155–158.
- [Ont] Government of Ontario. "French as a second language". Retrieved from: <http://www.edu.gov.on.ca/eng/amenagement/FLS.html>, November 2017.
- [OWC14] Otsuka, E.; Wallace, S.A.; Chiu, D. "Design and evaluation of a twitter hashtag recommendation system". In: 18th International Database Engineering & Applications Symposium, 2014, pp. 330–333.
- [PB07] Pazzani, J. M.; Billsus, D. "The Adaptive Web: Content-Based Recommendation Systems", . Springer Berlin Heidelberg, 2007, 341p.
- [QCDM⁺11] Quattrone, G.; Capra, L.; De Meo, P.; Ferrara, E.; Ursino, D. "Effective Retrieval of Resources in Folksonomies Using a New Tag Similarity Measure". In: 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 545–550.

- [QSCT⁺17] Santos, C.; Cunha, H.; Teixeira, C.; Souza, D.; Tietzmann, R.; Manssour, I.; Silveira, S. M.; Träsel, M.; Ruiz, A. R. D.; Barros, R. "Media Professionals Opinions About Interactive Visualizations of Political Polarization During Brazilian Presidential Campaigns on Twitter". In: 50th Hawaii International Conference on System Sciences, 2017, pp. 1891 – 1900.
- [RAC⁺02] Rashid, A.; Albert, I.; Cosley, D.; Lam, S. K.; McNee, M. S.; Konstan, A. J.; Riedl, J. "Getting to know you: learning new user preferences in recommender systems". In: 7th International Conference on Intelligent User Interfaces, 2002, pp. 127–134.
- [Rap02] Rapp, R. "The computation of word associations: Comparing syntagmatic and paradigmatic approaches". In: 19th International Conference on Computational Linguistics, 2002, pp. 1–7.
- [RGH⁺14] Ronen, S.; Gonçalves, B.; Hu, K.; Vespignani, A.; Pinker, S.; Hidalgo, C. "Links that speak: The global language network and its association with global fame". *The National Academy of Sciences of the United States of America*, vol. 111-52, Dec 2014, pp. E5616–E5622.
- [SB00] Santos, D.; Bick, E. "Providing Internet Access to Portuguese Corpora: the AC/DC Project". LREC, 2000, 85p.
- [Seg07] Segaran, T. "Programming collective intelligence: building smart web 2.0 applications". O'Reilly Media, 2007, 368p.
- [SGMB08] Shepitsen, A.; Gemmell, J.; Mobasher, B.; Burke, R. "Personalized recommendation in social tagging systems using hierarchical clustering". In: 1st ACM Conference on Recommender Systems, 2008, pp. 259–266.
- [SGP11] Stiller, J.; Gäde, M.; Petras, V. "Is tagging multilingual?: A case study with bibsonomy". In: 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, 2011, pp. 421–422.
- [SH09] Segaran, T.; Hammerbacher, J. "Beautiful data: the stories behind elegant data solutions". O'Reilly Media, 2009, 386p.
- [SK09] Su, X.; Khoshgoftaar, T. "A survey of collaborative filtering techniques". *Advances in Artificial Intelligence*, vol. 4, Jan 2009, pp. 2–4.
- [SKK10] Strohmaier, M.; Körner, C.; Kern, R. "Why do users tag? detecting users' motivation for tagging in social tagging systems". In: 4th International AAAI Conference on Weblogs and Social Media, 2010, pp. 339 – 342.

- [SKK12] Strohmaier, M.; Körner, C.; Kern, R. "Understanding why users tag: A survey of tagging motivation literature and results from an empirical study". *Web Semantics*, vol 17, Dec 2012, pp. 1–11, .
- [SLR⁺06] Sen, S.; Lam, S.; Rashid, A. M.; Cosley, D.; Frankowski, D.; Osterhouse, J.; Harper, F. M.; Riedl, J. "Tagging, communities, vocabulary, evolution". In: 20th Anniversary Conference on Computer Supported Cooperative Work, 2006, pp. 181–190.
- [SNPAR14] Santos-Neto, E.; Pontes, T.; Almeida, J.; Ripeanu. M. "On the choice of data sources to improve content discoverability via textual feature optimization". In: 25th ACM Conference on Hypertext and Social Media, 2014, pp. 273–278.
- [Sol16] Solon, O. "What's next for flickr after yahoo's sale?". Retrieved from: <http://google/2RxxBCJu1>, July 2016.
- [SS16] Sedhai S.; Sun, A. "Effect of spam on hashtag recommendation for tweets". In: 25th International Conference Companion on World Wide Web, 2016, pp. 97–98.
- [SvZ08] Sigurbjornsson, B.; Zwol, R. "Flickr tag recommendation based on collective knowledge". In: 17th International Conference on World Wide Web, 2008, pp. 327–336.
- [Tur05] Turkle, S. "The second self: Computers and the human spirit". Mit Press, 2005, 387p.
- [Use17] UserTesting. "The 2017 UX and User Research Industry Survey Results Are In!". Retrieved from: <http://www.usertesting.com/blog/2017/01/30/2017-ux-and-user-research-industry-survey-results>, March 2017.
- [Ver06] Veres, C. "The language of folksonomies: What tags reveal about user classification". In: International Conference on Application of Natural Language to Information Systems, 2006, pp. 58–69.
- [VSR09] Vig, J.; Sen, S.; Riedl, J. "Tagsplanations: Explaining recommendations using tags". In: 4th International Conference on Intelligent User Interfaces, 2009, pp. 47–56.
- [Z⁺12] Ziesemer, de C. A. A. "Recomendação de tags para mídia social colaborativa: da generalização à personalização", Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, 2012, 106p.
- [ZBS16a] Ziesemer, A.; Blustein, J.; Silveira, M. "Users tagging behavior and the effect of recommendation". In: 15th Brazilian Symposium on Human Factors in Computer Systems, 2016, pp. 1–4.
- [ZBS16b] Ziesemer, A.; Blustein, J.; Silveira, S. M. "Multilingual tagging behaviour: The role of recommender systems". In: 24th Conference on User Modeling Adaptation and Personalization, 2016, 2p.

- [ZMS] Ziesemer, A.; Müller, L.; Silveira, S. M. “More than content classification: Self-expression through image tagging”. In: 15th International Conference WWW/Internet, 2016, pp. 35-42.
- [ZNHP17] Zhang, Y.; Ni, M.; Han, W.; Pang, J. “Does #like4like indeed provoke more likes?”. In: 1st International Conference on Web Intelligence, 2017, pp. 179–186.
- [ZO11] Ziesemer, A.; Oliveira, J. B. S. “How to know what do you want? a survey of recommender systems and the next generation”. In: 8th Brazilian Symposium on Collaborative Systems, 2011 pp. 104–111.

A. Appendix A

A.1 PUCRS Approval Letter

A.2 Dalhousie University Approval Letter



Pontifícia Universidade Católica do Rio Grande do Sul
PRÓ-REITORIA DE PESQUISA, INOVAÇÃO E DESENVOLVIMENTO
COMITÊ DE ÉTICA EM PESQUISA

Porto Alegre, march 10, 2016.

Research Ethic Committee Approval Letter

The Research Ethics Committee evaluated the research Project:
“Predicting User Behavior on Tag Recommender Systems Based on Cultural Patterns of Communication (portuguese: Predição de Comportamento de Usuários em Sistemas de Recomendação de Tags com Base em Padrões Culturais de Comunicação)” under the protocol **CAAE: 49282415.0.0000.5336** with **Milene Selbach Silveira** and **Angelina Zieseimer**.

The REC-PUCRS reached a final approval on Jan 12nd 2016.

Prof. Dr. Rodolfo Herberto Schneider

Coordinator of Research Ethic Committee

Pontifical Catholic University of Rio Grande do Sul

PUCRS | **Campus Central**
Av. Ipiranga, 6681 – 5ºandar – CEP: 90619-900
Sala 505 – Fone Fax: (51) 3320-3345
E-mail: cep@pucrs.br

Figure A.1: Letter of approval for conducting research with humans at PUCRS – Brazil.



**Social Sciences & Humanities Research Ethics Board
Letter of Approval**

July 07, 2016

Angelina Zieseemer
Computer Science/Computer Science

Dear Angelina,

REB #: 2016-3848
Project Title: Predicting User Behavior on Tag Recommender Systems Based on Cultural Patterns of Communication
Effective Date: July 07, 2016
Expiry Date: July 07, 2017

The Social Sciences & Humanities Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans*. This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Sincerely,

Dr. Karen Beazley, Chair

Figure A.2: Letter of approval for conducting research with humans at Dalhousie University – Canada.

B. Appendix B

B.1 Consent Form Dalhousie University

CONSENT FORM

Project title: Predicting User Behavior on Tag Recommender Systems Based on Cultural Patterns of Communication.

Lead researchers: Angelina Ziesemer, Dalhousie University, an492994@dal.ca

Other researchers: James Blustein, Dalhousie University, jamie@cs.dal.ca

Funding provided by: CNPq

Introduction

We invite you to take part in this research study being conducted by me, (Angelina Ziesemer, M.Sc), a graduate student in the Department of Computer Science at Dalhousie University as part of my doctoral degree program. The purpose of this research is to investigate the usage of tags and recommender systems by distinct cultures. The tags you will assigned will be compared with tags from Brazilians participants. We aim to learn how people from distinct cultures assign tags to photos using two different interfaces (one of them having a support of a recommender system to suggest tags) and if the structure of tags change among cultures. To be eligible to participate in this study you must be Canadian (citizen or permanent resident), have used social networks, and be at least 18 years old.

If you volunteer as participant in this research you will be asked to used two distinct interfaces to assign tags to photos and answer a survey. In the first interface (A) you will be asked to assign at least four tags to four photos; in the second interface (B) you will be asked to assign tags to 7 photos. After, you will be asked to complete a survey (C) about your tag habits, social networks you have used, cultural background and preferences with respect to the type of language to tag. This study will take no more than 30 minutes. Responses will be stored by the Survey Monkey cloud based service. Survey Monkey hosted his data server outside of Canada (United of States and Ireland). According to the his privacy policy, SurveyMonkey merely acts as a custodian on behalf of the survey creator (the researcher conducting this study) who controls your data, your survey responses are owned and managed by the survey creator. We keep this survey protected by password. Also, the tags collected during the recommendation stage will be kept in a server in Brazil at PUCRS (Pontifical Catholic University of Rio Grande do Sul). Moreover, no sensitive personal information will be gathered, so your identity will not be revealed in any report.

Taking part in this research is entirely your choice. You will be compensated \$10.00 for your time. You can end your participation at any time during the study, as soon you asked me to remove your data before the period of participation is finished. The risks associated with this study are no greater than those you encounter in your everyday life.

Although your participation is not anonymous, your data will be anonymized in all reports and presentations. The individual code provide by us to start the tasks will be used to link your data among the three stages/interfaces of the study. All the data you provide will be securely stored. All records from the recommender system will be kept in a server database protected by password and data gathered by the interface without recommendation and survey will be kept in a cloud-based service protect by password as well.

Results will be presented aggregate and tags will be explicitly presented but not linked with any data that could identify you. Information that you provide to us will be kept private. Only the research team will have access to this information. We will describe and share our findings in conference papers, theses and journals. You can obtain these results by including your e-mail contact at the end of the signature.

This research is funding by the Brazilian Government, CNPq - Science Without Borders Program. Student scholarship award number: 201712/2015-6

You should discuss any questions you have about this study with me Angelina Ziesemer or if you have questions later, please contact me by: an492994@dal.ca

If you have any ethical concerns about your participation in this research, you may also contact Research Ethics, Dalhousie University at (902) 494-1462, or email: ethics@dal.ca (and reference REB file #2016.3848).

B.2 Consent Form PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul Faculdade de Informática Predicting User Behavior on Tag Recommender Systems Based on Cultural Patterns of Communication.

You are invited to participate in an academic research entitled “Predicting User Behavior on Tag Recommender Systems Based on Cultural Patterns of Communication”, which aims to analyze aspects of human-computer interaction, conceptual or software related, from the point of view of end users (actual or potential), expert users in Human-Computer Interaction, and/or experts in the field of the application. Please note that the aim of this study is the analysis of the aspects of human-computer interaction, and not the participant’s expertise in this field.

For the data collection we may be used different techniques, such as online surveys, log analysis of online tools provided by the researchers and also public information available on social networks.

All personal information resulting from this research will be treated confidentially. We also highlight that:

Anonymity must be preserved in any document published in scientific forums (such as conference papers, journals, books and similar reports) or educational (such as handouts courses, slides, etc.). At any time during the study, participant can withdraw his/her consent, and it will not bring any consequences for him/her. Moreover, your data will be removed from this study. Participants who are minors must obligatorily to present the consent of their legal representative to be able to

participate in this study, which will be declared aware of the study to be carried out by signature in this Consent Form. The research team is entitled to use the data collected, maintained the above conditions, for academic purposes, educational and/or analysis, development and evaluation systems.

If you have any doubt please contact the FACIN, PUCRS - Avenida Ipiranga, 6681 - Prédio 32 - 90619-900 - Porto Alegre - RS. Tel: + 55 (51) 3320-3558. Phone number on non business hours: +55 51 91868877 Regarding any ethical concerns about your participation in this research, you can also contact the Ethics Research Committee (Comitê de ética em Pesquisa - CEP) da PUCRS. Av. Ipiranga 6681, Prédio 40 - Sala 505. Porto Alegre, RS - Brasil - CEP: 90619-900. From Monday to Friday. Hours: 8h30 to 12h; 13h30min to 17h. Phone: +55 (51) 3320-3345.

C. Appendix C

C.1 Images

<i>P_a</i>	https://www.mensagenscomamor.com/images/jpgs/img/i/imagens_de_gatos_fofos1_14.jpg
<i>P_b</i>	https://goo.gl/8cq3bD
<i>P_c</i>	https://tinyurl.com/yb4dulk4
<i>P_d</i>	https://southernmorning.files.wordpress.com/2012/04/beach-chairs-3.jpg?w=440&h=300&crop=1
<i>P_e</i>	https://goo.gl/UQ6YFP
<i>P_f</i>	https://rareyal fresco.wordpress.com/2013/01/24/as-parisian-as-it-gets/
<i>P_g</i>	http://1.bp.blogspot.com/-vi5LORWr79s/UveIr1ZSJXI/AAAAAAAAp gg/-INUkuyRdG8/s1600/11_toddler-naps-with-puppy-theo-and-beau-2-9.jpg

Table C.1: URL locations for the images used during the user studies.