

Number 875



**UNIVERSITY OF
CAMBRIDGE**

Computer Laboratory

The language of collaborative tagging

Theodosia Togia

September 2015

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500
<http://www.cl.cam.ac.uk/>

© 2015 Theodosia Togia

This technical report is based on a dissertation submitted September 2015 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Trinity Hall.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

<http://www.cl.cam.ac.uk/techreports/>

ISSN 1476-2986

Abstract

Collaborative tagging is the process whereby people attach keywords, known as tags, to digital resources, such as text and images, in order to render them retrievable in the future. This thesis investigates how tags submitted by users in collaborative tagging systems function as descriptors of a resource's perceived content. Using computational and theoretical tools, I compare collaborative tagging with natural language description in order to determine whether or to what extent the former behaves as the latter.

I start the investigation by collecting a corpus of tagged images and exploring the relationship between a resource and a tag using theories from different disciplines, such as Library Science, Semiotics and Information Retrieval. Then, I study the lexical characteristics of individual tags, suggesting that tags behave as natural language words. The next step is to examine how tags combine when annotating a resource. It will be shown that similar combinatorial constraints hold for tags assigned to a resource and for words as used in coherent text. This realisation will lead to the question of whether the similar combinatorial patterns between tags and words are due to implicit semantic relations between the tags. To provide an answer, I conduct an experiment asking humans to submit both tags and textual descriptions for a set of images, constructing a parallel corpus of more than one thousand tags-text annotations. Analysis of this parallel corpus provides evidence that a large number of tag pairs are connected via implicit semantic relations, whose nature is described. Finally, I investigate whether it is possible to automatically identify the semantically related tag pairs and make explicit their relationship, even in the absence of supporting image-specific text. I construct and evaluate a proof-of-concept system to demonstrate that this task is attainable.

Acknowledgements

I feel privileged to have been supervised by Ann Copestake, who patiently guided me through this PhD, shaping the way I do research while giving me freedom to pursue my own ideas. I can't thank her enough for offering me regular meetings and helping me keep on track with my schedule. Her careful reading of drafts and reliable advice have been instrumental in the completion of this thesis. I am also immensely grateful to my examiners, Simone Teufel and Mark Stevenson for their thorough reading of my thesis, their constructive comments and a surprisingly enjoyable three-hour viva. Additionally, I would like to thank Steve Clark for giving me crucial feedback during the first two years of my PhD as well as my friends from the Computer Lab, with whom I have had stimulating discussions.

A very warm and special thanks goes to Daniel Gutiérrez Trápaga, for being by my side, supporting me through difficult times and making every minute of this journey worth experiencing. I am also grateful to my friends from Trinity Hall who participated in pilot experiments that I conducted in the course of this PhD.

This thesis would have been impossible without the love and support I received from my family. My siblings, Ioanna, Giannis and Niki, with whom I have shared dreams, successes, worries and many beautiful moments, have been a constant source of encouragement. My mother, Angeliki, who monitored my progress throughout my school years, has been a role-model of courage and devotion, that has helped me overcome numerous challenges in life. My father, Sokratis, who taught me how to solve equations at the age of ten and always encouraged me to 'think big', has been a source of inspiration ever since I can remember. Μαμά και μπαμπά, τα κατάφερα! Αυτή η διατριβή είναι αφιερωμένη σε σας.

Contents

1	Introduction	11
1.1	Tagging	11
1.2	Folksonomy	13
1.3	Tag clouds	15
1.4	Thesis outline	16
1.5	Summary	17
2	Image tagging	19
2.1	Corpus of tagged images	19
2.2	The nature of digital images	21
2.3	Tags as metadata	25
2.3.1	Subject of a document	26
2.3.2	Theme and rheme	26
2.3.3	Ofness and aboutness	27
2.3.4	Pre-iconography, iconography and iconology	28
2.4	Summary	30
3	Tags as words	31
3.1	Probability distribution of tags	31
3.1.1	Head and tail of distribution	34
3.1.2	Resource-specific tag distribution	37
3.2	Tag occurrences in natural language text	38
3.3	Tags and parts of speech	39
3.4	Tags and word categories	40
3.5	Summary	41
4	Tags in combinations	43
4.1	Relationships between words	43
4.1.1	Syntagmatic and paradigmatic relations	43
4.1.2	Lexical cohesion	45
4.2	Tag co-occurrence patterns	47
4.2.1	Co-occurrence vectors	47
4.2.2	Syntagmatic relations in folksonomy and natural language	52
4.3	Tag similarity	56
4.3.1	Similarity vectors	56
4.3.2	Paradigmatic relations in folksonomy and natural language	56
4.4	Summary	59

5	A parallel corpus of tags and text	61
5.1	Objectives	61
5.2	Experimental design	62
5.2.1	Variables and conditions	62
5.2.2	Experimental and control group	63
5.2.3	Stimuli	63
5.3	Piloting	64
5.3.1	Pilot study 1 (October 2012)	64
5.3.2	Pilot study 2 (November 2012)	66
5.3.3	Pilot study 3 (January 2013)	67
5.4	Final Experiment	68
5.4.1	Recruitment and data collection	68
5.4.2	Task presentation	68
5.4.3	Interface	70
5.5	Resulting corpus	72
5.6	Existing corpora	76
5.7	Summary	77
6	Implicit inter-tag relations	79
6.1	Distributions of tags and words	79
6.2	Tag-word overlaps	82
6.3	Inter-tag relations in parallel corpus	84
6.3.1	Path constraints	85
6.3.2	Pattern constraints	89
6.3.3	Overlaps and compositionality	91
6.4	Related work	92
6.4.1	Pre-specified relations	92
6.4.2	Open-ended relations	93
6.4.3	Relation extraction in folksonomies	95
6.5	Summary	95
7	Postulating tag-relation-tag triples	97
7.1	Task description	97
7.1.1	Overview	97
7.1.2	Feasibility	100
7.2	Producing triples	102
7.2.1	Well-formedness constraints	103
7.2.2	Plausibility constraints	109
7.3	Using specialised text corpora	113
7.3.1	Image caption corpus	113
7.3.2	Visual arts corpus	114
7.3.3	Wikipedia vs. specialised corpora	115
7.4	Summary	118
8	Evaluating postulated triples	119
8.1	Measuring the quality of suggested triples	119
8.2	Baseline	123
8.3	Testbeds	124

8.3.1	Unseen parallel subcorpus	124
8.3.2	A posteriori human judgements	126
8.4	Pilot experiments	128
8.4.1	Pilot experiment 1	128
8.4.2	Pilot experiment 2	131
8.4.3	Pilot experiment 3	132
8.5	Main experiment	135
8.5.1	Process	136
8.5.2	Inter-rater reliability	138
8.5.3	System performance	140
8.5.4	Phrases suggested by participants	143
8.6	Summary	146
9	Conclusions	147
9.1	Contributions of the thesis	147
9.2	Further work	148
	References	150
A	Parallel corpus experiments	163
A.1	Pilot 1	163
A.2	Pilot 2	166
A.3	Pilot 3	167
A.4	Final experiment	169
A.4.1	Recruitment Email	169
A.4.2	Differentiating image order	170
A.4.3	Phase One interface	172
A.4.4	Phase Two interface	180
B	Evaluation experiments	183
B.1	Pilot experiment 1 (with binary judgements)	183
B.1.1	Screenshots	183
B.1.2	Comments	184
B.1.3	Responses	185
B.2	Pilot experiment 2 (testing instructions)	188
B.3	Pilot experiment 3 (measuring inter-rater reliability)	190
B.3.1	Interface	190
B.3.2	Responses	191
B.4	Main experiment	193
B.4.1	Screenshot	193
B.4.2	Scores per image per system	194
B.4.3	Phrases suggested by participants	197

Acronyms

ADP	Abstract Dependency Pattern
ANOVA	Analysis of Variance
AUC	Area Under Curve
IDP	Instantiated Dependency Pattern
IE	Information Extraction
IR	Information Retrieval
IRR	Inter-Rater Reliability
LIS	Library and Information Science
MRS	Minimal Recursion Semantics
MWT	Multi-Word Tag
NE	Named Entities
NLP	Natural Language Processing
PMI	Pointwise Mutual Information
POS	Part of Speech
RTC	Resource-based Tag-tag Co-occurrence

Chapter 1

Introduction

In online collaborative tagging platforms, users organise digital resources by labelling them with keywords, known as *tags*.¹ Used to facilitate later retrieval or discovery of a resource, tags are the vehicle by which parts of the resource’s content (such as described entities, events, concepts) or other associated information (such as style, creator and so on) are isolated and recorded. A set of tags can, thus, be used to provide a rudimentary description of a resource.

Describing by tagging can be compared to a linguistic process, in that tags tend to be equivalent to natural language words or phrases, but a process without any obvious combinatory rules between individual tags. In this thesis, I explore how tagging, a less structured analogue to natural language, attempts to capture and communicate the meaning of a document. More specifically, I will aim to provide evidence that:

1. Tags behave like words, both **a)** in *isolation* and **b)** in *combinations*.
2. Tags can compose to assign complex meanings via *implicit semantic relations* that underlie particular tag pairs labelling the same digital resource.
3. It is possible to automatically **a)** *detect* which tags annotating the same document are semantically connected and **b)** *postulate* acceptable explicit representations of their implicit relations.

For the purposes of this research, I have conducted experiments on tagged *images*, as opposed to digital resources that contain language (e.g. webpages, videos or audio). The main reason is that, when labelling images, users perform tagging at its purest form, without the interference of language already existing in the resource.

A further contribution of the research presented here, in addition to providing evidence for the above claims, is the creation of a *parallel corpus* of tags and textual descriptions with respect to images. The corpus consists of 1,090 parallel annotations, collected through an experiment with humans (see Chapter 5).

1.1 Tagging

Tagging has been described in the literature as annotation with *uncontrolled vocabulary* (Mathes, 2004; Weller, 2007; Shepitsen et al., 2008; Heymann and Garcia-Molina, 2009;

¹Online digital resources are documents, such as images, videos or text accessible on the World Wide Web. From now on the terms ‘resource’ and ‘document’ will be used interchangeably.

Trant, 2009a). It utilises a *vocabulary*, that is isolated terms, because syntactic structure is not possible between the tags. This vocabulary is *uncontrolled* because there are no lexical boundaries imposed, allowing for misspellings, non-words, synonymy, polysemy, redundancy and contradiction. Tagging can be seen as an alternative to subject cataloguing, typically used in libraries and museums. In both cases, authors provide keywords for documents for the purpose of facilitating retrieval and collocating similar documents. However, tagging is more liberal than the ‘prescriptive’ professional cataloguing (Tennis, 2006), as it does not follow any pre-determined classification scheme.

Tagging evolved as a practice within desktop indexing systems of the 1980s (e.g. Lotus Magellan) that allowed user-generated keywords to enter the index. In the 1990s the computer communication service Compuserve allowed users to add their own keywords to documents that they submitted to the network (Vander-Wal, 2007). With the advent of Web 2.0 (O’Reilly, 2007) in the early 2000s, more and more ordinary internet users became creators, as opposed to mere consumers, of content. This development gave rise to sites like Bitzi², where users could share and comment on documents that they had annotated with free-form keywords. The first important online tagging system was Delicious³, a bookmark organisation system created in 2003, that allowed registered users to label bookmarks (i.e. favourite webpages) contributed by themselves or by other users. The addition of user registration, and, hence, user identity in the organisation system created a tri-partite network of users, tags and digital resources, which can reveal interesting tagging behaviour patterns (see §1.2 for a discussion). With respect to images, the first large-scale tagging website was Flickr⁴, which allowed pictures to be annotated by their owners who wished to retrieve them later or render them searchable by the general public. Flickr differed from Delicious in that it only allowed a resource to be tagged by its author, as opposed to everyone. The possibility of any user tagging any resource, which was adopted by Delicious, is especially interesting, since it results in a bag of tags associated with each document, allowing for the construction of the document’s ‘social meaning’ (see §1.2). These two systems “were causing quite a stir on many of the information science list serves as the tagging seemed to be working for finding things” (Vander-Wal, 2007). Delicious and Flickr laid the foundations for numerous online tagging systems that have been developed since then.

Strohmaier et al. (2010) distinguish between two types of activities that users perform on a tagging system, *categorising* and *describing* resources. When categorising, users provide tags that aim to classify a document under high-level categories (e.g. ‘fashion’, ‘education’, ‘design’ etc.). According to the author, Flickr images tend to be tagged in this way since the platform’s interface prioritises *browsing* over search as a means for discovering images, hence encouraging users to provide tags that correspond to browsable categories. On the contrary, when describing a resource, users generate tags that resemble “games with a purpose”, also known as ESP (“Everyone Should Play”) games (von Ahn, 2006), in which users try to guess each other’s tags for an image, which encourages them to label the image with highly descriptive tags (e.g. ‘chair’, ‘people’, ‘sitting’, ‘classroom’, ‘lesson’ etc). Describing a resource with tags aims to facilitate not later browsing but later *searching*. The distinction between browsing (“exploring a problem space to formulate questions”) and searching (“looking for answers to specifically formulated questions”)

²<http://web.archive.org/web/20131229171319/http://bitzi.com/>

³<https://delicious.com> (originally del.icio.us)

⁴<https://www.flickr.com/>

(Mathes, 2004) in the tagging process is very important, as these two different goals can motivate users to submit different kinds of tags. Strohmaier et al. (2010) show that tagging motivation is also highly influenced by the idiosyncrasies of the tagging system. For instance, as mentioned, a tagging platform with a well-designed browsing interface might motivate users to tag with categories in mind. Alternatively, a system that focuses on search might elicit more descriptive tags. This thesis deals with tagging intended for description rather than categorisation of documents.

1.2 Folksonomy

Online tagging systems are typically social environments, allowing a large number of users to register and collectively organise digital resources. This kind of process, known as *collaborative tagging*, is open, liberal and decentralised. Yet, if performed for a period of time by a large number of users, tagging results in a complex system; structure emerges out of uncontrolled vocabulary as users reach an agreement on what keywords to use (Halpin et al., 2007), either for particular documents or across the system (further discussion in chapter 3). This structure is called *folksonomy* (Vander-Wal, 2007) because it resembles a ‘folk’ (bottom-up and flat) ‘taxonomy’ (organisation) of objects.

The term ‘folksonomy’ was coined by Thomas Vander-Wal, who explains its characteristics in his widely cited blog post “Folksonomy Coinage and Definition” (2007). The author describes folksonomy as the result of “personal free tagging [...] for one’s own retrieval [...] done in a social environment”. Users typically share their tags with the rest of the community.

Folksonomy has been compared to taxonomy (Shirky, 2005), where objects belong to hierarchically organised categories; but there are two main differences: **i**) folksonomy offers bottom-up and flat categorisation; for example, a particular digital resource can belong to as many ‘categories’ as the tags assigned to it, **ii**) there is no explicit relation, such as ‘is-a’ or ‘instance-of’, between the category and the digital object. Folksonomy has also been referred to as *ethnolocation* (Merholz, 2004), but as Mathes (2004) comments, this is an unsuccessful term since there is nothing about classification in collaborative tagging; classification schemes organise an item under a single category with explicit relations, but tagging systems offer just organisation, which is less rigorous and less restrictive.

A few years before his complete definition of folksonomy, Vander-Wal (2005) had distinguished between *broad* and *narrow* folksonomies. The former are created when multiple users are allowed to annotate the same resource (e.g. in Delicious, different users can label the same bookmarked webpage), resulting in some tags gaining popularity over others. The latter are created when only the author of a particular resource is allowed to add tags to it (e.g. in Flickr images; see §1.1). Broad folksonomies are the focus of this thesis because the tendencies revealed from multiple people annotating the same image can offer a valuable insight into the public understanding of a document.

The following formal definition, which sees folksonomy as a *triadic formal context* (Lehmann and Wille, 1995), is widely cited in the folksonomy literature; presented here from (Hotho et al., 2006) with slight notational adaptations:

Definition 1 *A folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$, where U , T and R are finite sets of users, tags and resources respectively; Y is a ternary relation that holds between them (called ‘tag assignment’), $Y \subseteq U \times T \times R$.*

The ternary relationship Y can have different constraints depending on the tagging platform. As mentioned, some systems allow any resource to be annotated by any users with any number of tags; some allow one user per resource; some might impose restrictions on the number of tags per resource.

Folksonomy has also been described as a tri-partite undirected hypergraph⁵ with three types of nodes (users, tags and resources) and hyperedges connecting them. Conceptualising folksonomy as a graph allows its analysis with graph-based techniques (e.g. clustering tags, users or resources using centrality measures). The definition below (Mika (2005)) is quoted from Schmitz et al. (2006) with some adaptations:

Definition 2 A hypergraph of a folksonomy $\mathbb{F} := (U, T, R, Y)$ (as per the previous definition) is a simple tripartite hypergraph $H(\mathbb{F}) = \langle V, E \rangle$, where H is the hypergraph, V are the vertices, $V = U \cup T \cup R$ and E are the edges, $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$.

Figure 1.1 illustrates the notion of folksonomy, while Figure 1.2 is a visualisation of a folksonomy as a tri-partite graph.

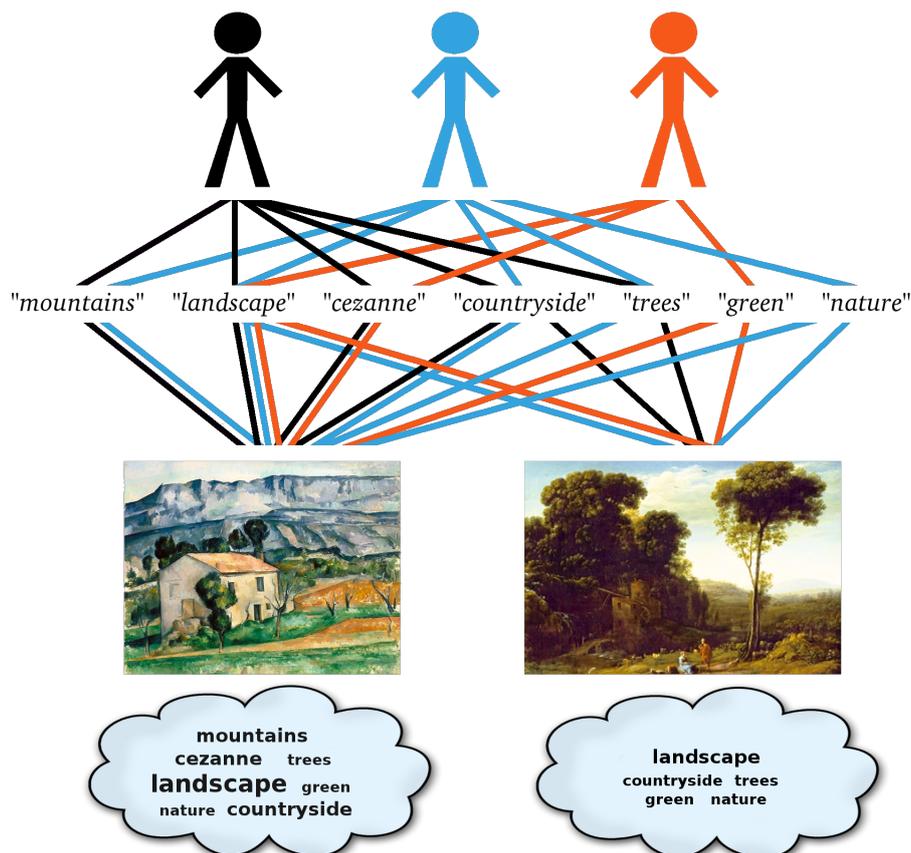


Figure 1.1: Folksonomy, comprising (from top to bottom) users, tags and resources. Below each image are the aggregated tag annotations from all users.

⁵A hypergraph is a generalised graph whose edges ('hyperedges') can connect more than two nodes at the same time.

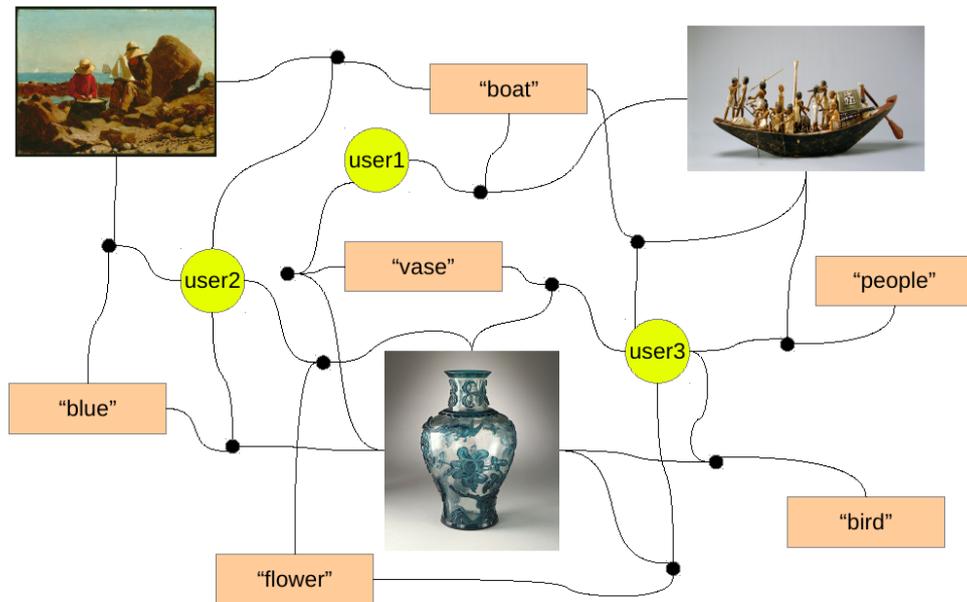


Figure 1.2: Folksonomy as a tri-partite graph. Black dots are meant to represent not nodes but connectors between three nodes simultaneously (an image, a user and a tag).

Vander-Wal argues that the tri-partite nature of a folksonomy is of vital importance: “the three tenets of a folksonomy: 1) tag; 2) object being tagged; and 3) identity [i.e. users], are core to disambiguation of tag terms and provide for a rich understanding of the object being tagged.” (Vander-Wal, 2007). Folksonomy exemplifies how web 2.0 allows us to “harness collective intelligence” (O’Reilly, 2007), with collaborative tagging being “the most successful tool for distributed cognition” (Steels, 2006).

1.3 Tag clouds

In broad folksonomies, each resource can be labelled by more than one user. A popular tag, that is one that has been assigned to a resource by a large number of different users, can be seen as more representative of the resource than, say, an hapax legomenon, that is a tag that has been provided by only one user. Each resource is, then, associated with a multiset of tags; the number of times each tag is repeated in the multiset corresponds to the number of times it has been used as a keyword for the resource. This multiset can be visualised as a *tag cloud*, in which the relative size of a tag depends on its relative weight in the multiset. Tag clouds have been described as “visualizations of a semantic field”, which is a “set of concepts connected to a focus” (Marinchev, 2006).

In the rest of this thesis, I will be using the term ‘tag cloud’ or simply ‘cloud’ to refer to the multiset of tags assigned to an image by a group of users rather than specifically to the visualisation of this multiset.

Tag clouds as semantic units

Since a tag cloud is a collection of descriptors for an image, it would be reasonable to wonder whether it models the meaning of the image. In other words, can we treat a tag cloud as a semantic unit? A positive answer would be in tune with the *statistical semantics hypothesis*. The statistical semantics hypothesis states that “statistical patterns of human

word usage can be used to figure out what people mean” (Furnas et al 1983, cited in Turney and Pantel 2010). This hypothesis is a generic term for more specific hypotheses such as the bag of words hypothesis (e.g. indexing in Information Retrieval), the distributional hypothesis, the extended distributional hypothesis, and the latent relation hypothesis (Turney and Pantel, 2010). We could treat a tag cloud, which consists of crowd-sourced individual annotations, as a bag of words that approximates the meaning of the image.

The notion of crowd-sourced meaning is in accordance with a late Wittgensteinian (Wittgenstein, 1953) understanding of semantics, which, in the folksonomy literature, has been called *social semantics* (Halpin, 2009). In *Philosophical Investigations* (1953), Wittgenstein posits that meaning is equivalent to *use*: one can explicate the meaning of a word (or a larger linguistic unit) by means of its function within a cultural environment (‘form of life’). One can grasp the meaning of a word by being exposed to various contexts in which the word is used. We can extend this theory to accommodate the meaning of an image tagged by multiple users. If an image is regarded as a content-bearing entity, analogous to a word, sentence or piece of discourse, then the tags which annotate it can be seen as examples of the image’s usage. For instance, if ten users have labelled a given image with the tag ‘tree’, then it can be said that this image has been *used* ten times to exemplify the concept of tree. The usage statistics for a tagged image (i.e. what concepts it exemplifies in what proportions), visualised as a tag cloud, can provide an insight into the image’s meaning, in the same way that contexts of use for a word can exemplify its socially agreed meaning.

1.4 Thesis outline

The rest of the thesis is organised as follows:

- In Chapter 2, I provide a theoretical background on image tagging that will assist in understanding the rest of the thesis. In particular, I describe the tag corpus that will be used in this thesis and discuss issues related to the meaning of digital images and image metadata.
- In Chapter 3, I show that individual tags act like words (cf. claim 1a, page 11). I do this by investigating the distribution of tags across the dataset and comparing lexical characteristics in tags to those observed in natural language.
- In Chapter 4, I provide evidence that tags annotating a given image are governed by combinatorial restrictions, similar to those found in natural language (cf. claim 1b, page 11).
- To investigate whether the observed natural-language-like restrictions on tag combinations are due to underlying semantic relations, I compiled a parallel corpus of tags and accompanying text, which I describe in Chapter 5.
- In Chapter 6, I analyse the parallel corpus created and show that tags indeed compose via implicit semantic relations in order to jointly assign meaning to an image (cf. claim 2, page 11).
- In Chapter 7, I show how semantically connected tag pairs can be detected in a tag cloud (cf. claim 3a, page 11) and how explicit representations of their implicit

relations can be postulated (cf. claim 3b, page 11) through the use of text corpora that do not describe the image in question.

- In Chapter 8, I evaluate the relations suggested for a set of images, proving the concept that acceptable relations can be postulated (cf. claim 3, 11).
- Finally, in Chapter 9, I conclude the thesis and discuss topics for further research.

1.5 Summary

In this chapter, I presented the goal of this research, namely investigating the role of tags as descriptors of images, and outlined the claims made in the rest of the thesis. I introduced the concept of tagging, discussing its history and characteristics, provided a formal definition of folksonomy, distinguishing between broad and narrow folksonomies, and discussed the emergence of tag clouds in folksonomy and their potential to function as semantic units.

Chapter 2

Image tagging

Before proceeding to a detailed account of experiments performed within this research, it is necessary to provide a theoretical framework that will facilitate understanding of the rest of the work. Section 2.1 describes the corpus of tagged images on which the majority of the experiments in this research have been based. Section 2.2 explores the nature of digital images, the understanding of which will clarify the various relationships between an image and its tags and, ultimately, assist in examining whether or how tags relate to each other. Finally, Section 2.3 compares tags to traditional metadata as descriptors of images.

2.1 Corpus of tagged images

To perform the experiments necessary for this research, the first step was to obtain a corpus of tagged images. Such a corpus must be extracted from an image folksonomy and fulfil the following requirements:

1. The users must be *everyday* people, who can interpret an image with varying degrees of sophistication, rather than professional cataloguers.
2. The images must be labelled for the purpose of later retrieval, either personal or social, which has been shown to result in highly *descriptive* tags, in contrast to other motivations, such as later browsing. (see §1.1)
3. The folksonomy must be *broad* (see §1.2), meaning that a given image can be labelled by more than one individual, which allows for image-specific tag clouds to be created.
4. The images must have enough complexity and some degree of diversity in order to elicit a wide *variety* of tags.

The most suitable folksonomy I found given the above requirements was the ‘Steve’ folksonomy (Trant, 2009b), created by ordinary people who tagged images of art objects from 21 institutions, mainly museums, in the United States.¹ The Steve folksonomy

¹Institutions involved were, among others, the Metropolitan Museum of Art (New York), Indianapolis Museum of Art, San Francisco Museum of Modern Art, Archives, The Cleveland Museum of Art, Denver Art Museum, Guggenheim Museum, Los Angeles County Museum of Art, Minneapolis Institute of Arts, The Rubin Museum of Art & Museum Informatics, and Think Design (Chun et al., 2006; Trant and Wyman, 2006).

resulted from a 3-year project (2006-2009), which aimed to collect user-contributed metadata for art images, thereby “bridging the distance between the professional, curatorial language of art history and public perceptions reflected, for example, in the way that searches are made of public art resources” (Trant and Wyman, 2006). The ultimate goal was to increase access to the images by facilitating retrieval. For the purposes of the Steve project, an online tagging tool was created², which allowed users from all backgrounds to annotate the images. Users were directed to the website by volunteer requests on mailing lists and blogs, and through publicity generated by the press, including the New York Times (Pink, 2005).

Throughout the project, tagging participants were presented with either of four different interfaces: **i)** an image with both professional metadata (e.g. title, creator etc.) and previous users’ tags being visible, **ii)** an image with metadata but no previous tags visible, **iii)** an image with existing tags but no metadata and **iv)** an image without any tags or metadata. It was found that the presence of professional metadata had no effect on the usefulness or the novelty of tags created by a user, but the availability of existing tags was shown to assist users in producing more original tags, since “users tended not to duplicate tags shown with a work of art, and instead entered different terms” (Trant and Wyman, 2006). The website was maintained until recently, and as of July 2014, it had collected 552,108 unique tags on 98,092 resources by 8,346 users.

Steve corpus For the purposes of the experiments I created a dataset by crawling the Steve project website³ during a three-week period in October and November 2011. The dataset compiled consists of 33,948 resources (nearly one third of the images available online), 65,065 distinct tags (slightly above one eighth of the ones submitted online) and 447,532 tag tokens. User information (i.e. which user labelled which image with which tags) could not be retrieved by crawling, so the resulting dataset is not a complete folksonomy, which would require tags, resources and users as part of its structure. What was accessible, however, is the *number* of individuals who had used a particular tag for a given image. This information was enough for the purposes of this research, since it allowed the study of tag clouds associated with images.

Other image folksonomies were also considered for downloading, however, none of these could fulfil the requirements set at the beginning of this research. For instance, Artigo⁴ (Bry and Wieser, 2012; Wieser, C., Bry, F., Alexandre, B., Lagrange, 2013) is similar to Steve in that it is a broad folksonomy of art images, yet its tags are being collected by means of “games with a purpose” (von Ahn, 2006) (see §1.1), where users get rewarded for guessing each other’s tags in real time. Players of these games are not explicitly asked to provide tags that will be useful as future access points to images. Hence, Artigo was dispreferred over concerns that its tags can be adapted to winning the game and may not necessarily be descriptive of the image. Two more large-scale image folksonomies, Flickr⁵ and Instagram⁶ were dismissed as they are *narrow*, that is they only allow an image to be tagged by its original creator (see §1.2), so they would not provide any information on the public perception of a given image.

²<http://web.archive.org/web/20140701205526/http://tagger.steve.museum/>

³<http://web.archive.org/web/20140701205526/http://tagger.steve.museum/>

⁴<http://www.artigo.org/>

⁵<https://www.flickr.com/>

⁶<http://instagram.com/>

2.2 The nature of digital images

In order to gain an insight into how users select tags for a given image, it is important to explore the nature of a digital image as a content-bearing entity, which can reveal the complexity of image interpretation.

A digital image is a document. Traditionally, a document has been seen as “any base of materially fixed knowledge that can be used for consultation, study or proof” (Union Française des Organismes de Documentation, cited in Briet 1951).⁷ Under this definition, the document is determined by means of its *medium*, good examples of documents being books, reports, photographic prints and so on. But what is a digital image? Are copies of a digital image separate documents? If we were to preserve the traditional definition of document, we would define a digital image by the physical object that carries it, for example, ‘what is shown on a screen’. However, a screen may contain more than one image, or pixel areas that would not be considered documents under any intuitive definition. We can attempt to define an image as a file on a hard drive, encoded in bits. That would also be problematic since bits are contiguous on a hard drive, so defining an image file as ‘what is stored on a hard drive’ would not set the limits of what is an what is not an image file. Buckland (1998) argues that with the advent of digital technology, a document can no longer be defined through its medium (i.e. it is not distinguishable by its medium), and suggested a functional definition in the spirit of Otlet (1934) and Briet’s (1951) work, who argued that a document is whatever *functions* as a document, that is, whatever is worthy of preserving and registering; even a meteorite or an animal.

When a user of an online tagging system labels an image of a painting, is it the digital document (i.e. what is shown on a screen) or the physical document (i.e. the actual painting in the museum) which is worthy of tagging? Tags such as “impressionist” are about the painting while tags such as “high definition” are about the digital image depicting the painting.

To understand how a digital image can encompass other documents, we can explain the notion of ‘document’ in the framework of Shannon and Weaver’s (1948) ‘mathematical theory of communication’. As demonstrated in the simplified model in Figure 2.1, the theory treats communication as an act of transmitting a message from one point to another. An information source (e.g. a human), also known as ‘transmitter’, possesses some information that they wish to transmit, that is send as a ‘message’, to a destination (e.g. another human), also known as the ‘receiver’ through a, potentially noisy, channel.

We could extend Shannon and Weaver’s model to accommodate the notion of ‘storage channel’, which can act as the medium on which an encoded message is recorded (Figure 2.2). This message has the potential to *function as a document* if it becomes worthy of cataloguing (or tagging).

⁷translated from French: “toute base de connaissance, fixée matériellement, susceptible d’être utilisée pour consultation, étude ou preuve”. (Unless otherwise stated, all translations are my own.)

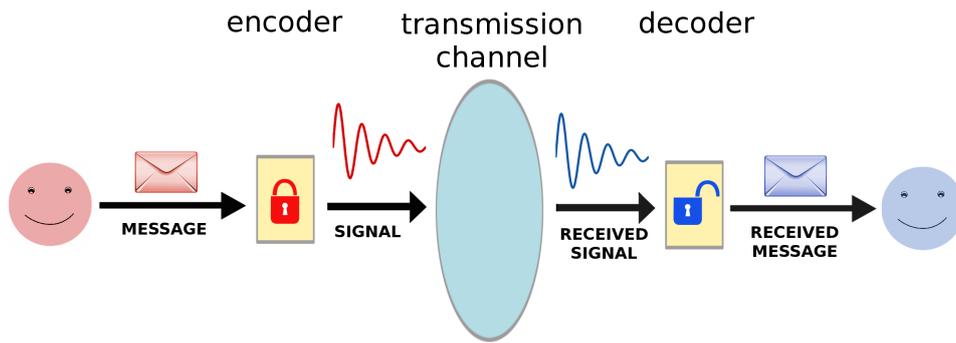


Figure 2.1: **Transmission channel:** A transmitter sends a piece of information (message), which is encoded as data (signal), passes through a transmission channel, is received as data (received signal) and gets decoded by the receiver into information (received message).

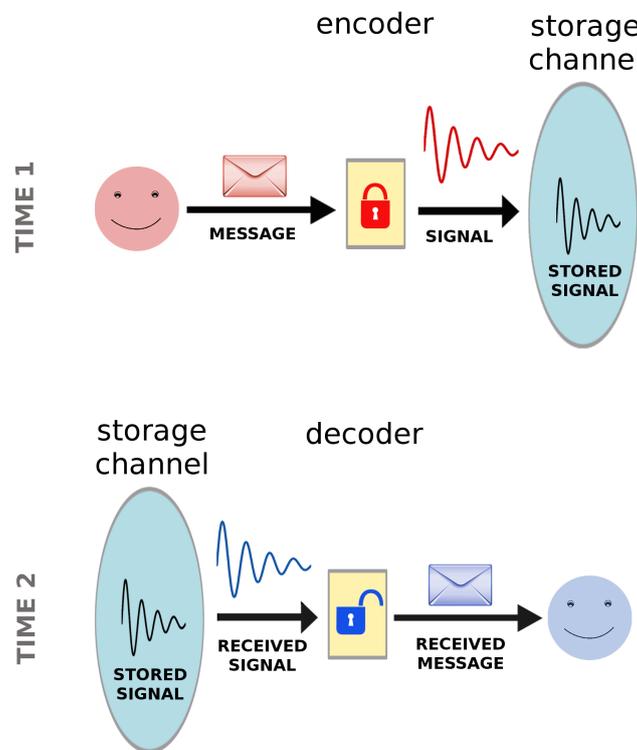


Figure 2.2: **Storage channel:** A transmitter sends a piece of information (message), which is encoded as data (signal), and gets stored into a storage channel. At any time in the future, one can receive data (received signal) from the storage channel and decode it into information (received message).

A user of a tagging platform can tag a painting, not by visiting the museum and attaching sticky notes to it, but indirectly, by means of a digital image which depicts it. However, this introduces some extra layer of decoding, which can manifest in the tagging process.⁸ As shown in Figure 2.3, at Time 1 the painter encodes his message in a

⁸It is important to clarify that the meaning attached to an image through tagging does not necessitate a user's (or different users') decoding of a single originally intended meaning. Users can manifest their

painting and at Time 2, this painting, in turn, becomes the message encoded as a digital image. Marshall McLuhan (1964) famously said “the medium is the message”, meaning, among other things, that what is the *medium* of one message (e.g. cardboard and oil paint encoding an image), is the *message* (content) of another medium (e.g. of the digital image). So, media can be nested and decoding them is a recursive process. A complicated example could be a digital image of a printed photograph showing people looking at a painting of a vase showing people, and so on.

Users of an image tagging system can attach keywords to any of the nested media which they may consider worthy of indexing, that is the ones to which they give a *document value*. The image in Figure 2.4c has been labelled with tags such as “box”, “bamboo”, “wood” and “vase”, which describe the content of the digital image (i.e. the wooden brush-pot), but also with tags such as “lady”, “farmer” and “tree”, which describe the visual content of the brush-pot (i.e. a scene). Similarly, the image in Figure 2.4a has been annotated with tags such as “table”, “chair” and “curtain” (i.e. content of digital image) but also with tags such as “single man” and “who is on tv” (content of television programme). Another example is the image in Figure 2.4b, which has tags such as “frame”, “gilt” and “gold” (i.e. content of digital image) but also tags such as “birthday”, “cake” and “table” (i.e. content of depicted painting).

personal understanding, or a generally agreed contemporary interpretation of an image, in the process of tagging. However, some decoding is necessary to prevent the image from being received as uninterpretable visual data; for instance, even recognising patterns, objects or culturally interesting features requires the knowledge of a code in Shannon and Weaver’s (1948) sense.

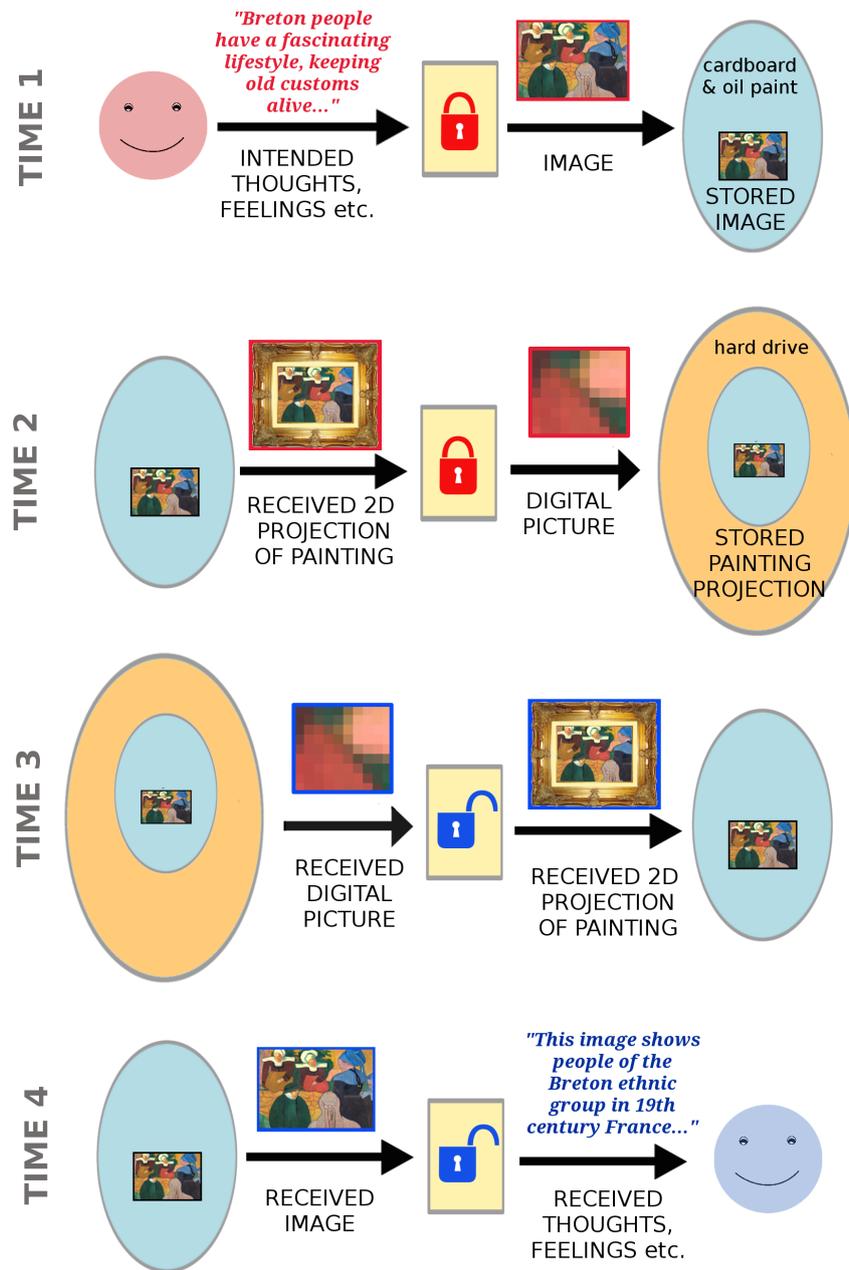


Figure 2.3: **Nested storage channels:** “Breton women at a wall” by Émile Bernard (1892); At Time 1, the painter encodes his thoughts as an image. At Time 2, a photographer encodes the (projection of the) painting into a photograph that gets stored as data into a hard drive. At Time 3, a user sees the digital image on the internet and decodes it from an array of pixels to the content (i.e. the painting). At Time 4, the user decodes the painting from visual data to a meaning.



1950s 1950sor60s 20thcentury aaaa absence asmallrestaurantempty
blackandwhite blackandwhitephotography breakfas
 deadplants **diner** diners diningtable domestic eating electricfan
 interior isthemanontvimportant **kitchen** kitchenfan light
 napkindispenser **napkinholder** napkins nopeople home
 photo **photograph** photography postwar **reflection**
 saltandpeppershakers silence singleman south stillife
 sunfilledwindow **sunlight** sunlightontable **table**
 tvdinner unnervng vvv wal wherearethekids whereisdadandmom
 astudy backlit bentwoodchairs whoisontv **window** cafe
fan glare highcontrast **chairs** curtain curtains cutainvalance
 emptiness lightblownoutatwindow lonely manontvinprofiletalking
 obliqueview oldschooltv oldtelevision openforbusiness pepper tidy **tv**
restaurant restaurantinterior salt saltandpepper strongcontrast
 study **studyrestaurantinterior** **television**
 whereisthefanpointto whereisthis whoisinthenextroom table&chairs

(a)



absentfather america amplephysiqueofwoman **birthday**
 manlurking
 bluepitcher boy buxom **cake** candle candlelight **candles**
 domesticscene **family** familygroup frame gilt glassware gold
 oil lampshade light lightonchildsface litcandleoncake littleboy
 manstandsinsadow mom mother motherandson motherservescake
red readdress redgown redroom seatedatatable servant seventhbirthday
 shadyman sharkchandelier slimman smallchild **table** whitelampshade
 shadows shadowy **birthdaycake** blueglasspitcher
 painting pitcher whitetablecloth woman
 overindulgence celebration **child** deepredwalls diningtable
 hiddenlightbehindman impressionism impressionistic interior
 makeawish manintheshadow oddcroppingofscene

(b)



18thcentury **asian bamboo** figures
chinese detailed everydayobjects farmer
 japaneselacqueredbox lady **landscape** hat
 people plants **rectangular** reed rocks
 squarecontainer storytellingbox stream sturdy tranquility
box boxlike bridge buildings carving **wood**
 fisherman fishermanwithox fishing garden gray
 man maroon narrated oriental ox **printed**
 scene scenery scenes scenic scholar **square**
 trees utilitarian **vase** vessel **water** woman
 trashbox wooden **black** tree intricate
 sandalwood

(c)

Figure 2.4: Nested media; images and tag clouds downloaded from Steve Tagger <http://web.archive.org/web/20140701205526/http://tagger.steve.museum/>

2.3 Tags as metadata

In the process of tagging, a user produces data, that is tags, related to a resource (here image), which itself is a form of data, since it contains encoded information. Therefore, tags are “data about data”, or *metadata*. According to the National Information Standards

Organization (2004), metadata are divided into three categories: **i)** structural, which specify the structure of the document (e.g. number of pages on a book), **ii)** administrative, which facilitate management of a document (e.g. file type, access permissions etc.) and **iii)** descriptive, which “describes a resource for purposes such as discovery and identification” and “can include elements such as title, abstract, author, and keywords”. Tags fall under the third category, since they are equivalent to keywords, that is “descriptive metadata which identifies and functions to organize information based on its intellectual content” (Mathes, 2004).

2.3.1 Subject of a document

In Information Retrieval (IR) and Library and Information Science (LIS) keywords (tags in our case) are known as *index terms* (or *subject indicators*), which Salton and Harman (2003) define as “content identifiers to information items and search requests”. Index terms “*individually* or in *combinations* are, supposedly, the names of subjects or topics” (Maron 1977; emphasis mine). In other words, they can work either alone or in groups in order to reveal what a document is about. Subjects of a document are equivalent to what has been called *topics* in the Text Retrieval Conference (TREC), which express information needs (Voorhees and Harman, 2005); one searches a collection of documents in order to fulfil their need for information on a particular topic.

A document can have a number of subjects with varying granularity. For example, the Magna Carta can be about human rights, democracy, fairness, limiting power abuse, King John’s obligation to obey the common law and so on. Deciding which subjects are associated with a document is crucial for determining what keywords one should use to index it for later retrieval. As Svenonius (2000) puts it, “Quality indexing, successful retrieval and effective automatic indexing depend on being able to define *subject*”, which, she says, can be defined “through the related concept of *aboutness*” (p.46, emphasis in original). Maron (1977) distinguishes between three types of aboutness: **i)** S-about (‘subjective about’), that reflects an annotator’s personal opinions about the document, **ii)** O-about (‘objective about’), that can be seen as what the document objectively is about, regardless of opinions and **iii)** R-about (‘retrieval about’), which can be represented by a probability distribution of index terms; the probability of each keyword is equivalent to the probability that someone would search for a document that ends up satisfying their information needs using that keyword. In Information Retrieval, R-about is represented by a bag of words (Salton et al., 1975), *automatically extracted* from the text in, or surrounding, a document (cf. ‘author aboutness’, Ingwersen 1992, p.50). In tagging systems, R-about is represented by a distribution of tags, which is an aggregation of subjective tags *manually assigned* by a group of users on a single document.

2.3.2 Theme and rheme

An important distinction with respect to the notion of subject in a document was made by Hutchins (1978). The author claims that index terms (i.e. names for subjects) attached to a document describe *presupposed* knowledge, as opposed to *new* knowledge, for someone searching a document collection. For instance, imagine a book that discusses Shakespeare’s plays and mentions, among other things, that some of his plays were influenced by those of Christopher Marlowe. This book is more likely to be useful to someone who is familiar with Shakespeare and wants to learn something new about him, than

to someone who is familiar with Marlowe and wants to expand their knowledge of the latter. In Hutchin’s approach, such a book would be about Shakespeare but not about Marlowe, because Shakespeare is the *theme* (the thing talked about) and Marlowe is part of the *rheme* (the details regarding the theme). Thus, subjects of a document are themes (presupposed knowledge) and not rhemes (new knowledge).

The idea of subjects as themes, in contrast to rhemes, can be extended to images. For instance, we can plausibly say that the image in Figure 2.5 tells the story of ‘two men’, of ‘a woman’, of ‘three people’, of ‘everyday life in early 20th century America’, of ‘old age co-existing with new age’, of ‘music’ and so on. However, it does not tell the story of ‘a woodfloor’, ‘a tablecloth’ or ‘sitting on a table’. We can distinguish between concepts that play a leading role in the image (themes) from concepts that are peripheral, providing context for the main concepts (rheme).⁹

Although professional keywords mainly reflect themes, user-contributed keywords (tags) can reflect both themes and rhemes. For example, the tags assigned to the image in Figure 2.5 refer not only obvious subjects such as ‘woman’, ‘men’ and ‘music’, but also to peripheral concepts such as ‘woodfloor’, ‘map’, ‘curtain’, ‘table’ and ‘flute’. One explanation for this behaviour can be that tags are more than just subject *indicators*. They can sometimes be subject *fragments*, that is tags that depend on other tags or even un-uttered concepts in order to signify a subject. For instance, the image in question does not tell the story of ‘a table’ but it might tell the story of ‘two men sitting in a table’.



Figure 2.5: “The Love Song” by Normal Rockwell (1926); downloaded from Steve Tagger <http://web.archive.org/web/20140701205526/http://tagger.steve.museum/>

2.3.3 Ofness and aboutness

So far we have seen that a document’s subject indicates what the document is about. In the context of images, however, *aboutness* is only one aspect of a subject, the other one being *ofness*. In an influential paper, Sara Shatford (1986) adopts a restricted definition of aboutness, which she defines as a relationship between an image and an abstract concept

⁹This distinction has been made for TREC images under the terms ‘foreground’ and ‘background’ concepts (Fujita, 2000).

(e.g. an image can be *about* happiness). For concrete objects seen in the image, the relationship is ofness (e.g. an image can be *of* a woman). The author suggests that archivists should aim to index images considering not only its *concrete and objective* topics (what the image is of) but also its *abstract and subjective* topics (what the image is about) given that information needs of a user can be for images of something or about something. For example, a photographer might be interested in retrieving images *of* a woman dancing flamenco, while an art historian might be interested in images *about* passion. They might both arrive at the same image, but it will be for different reasons and through different index terms. As Shatford comments, “the delight and frustration of pictorial resources is that a picture can mean different things to different people”. The ofness-aboutness dichotomy has been seen as equivalent to that of denotation and connotation (Yoon and O’Connor, 2010), terms often used in semiotic analysis of images (most notably in Barthes 1964). Krause (1988) describes concrete and abstract subjects of an image as ‘hard’ and ‘soft’ aspects of its content respectively.

2.3.4 Pre-iconography, iconography and iconology

The ofness and aboutness of an image is better understood in light of the image’s different levels of interpretation. When it comes to interpreting art, decoding a painting is often seen as a 3-level process, originally described by art historian Erwin Panofsky (1955): **i)** pre-iconography, **ii)** iconography and **iii)** iconology.

At the *pre-iconographic* level, one interprets an image with regard to its “primary or natural subject matter”, which can be factual (objects and actions, such as a woman, a table, people running, a thunder etc.) or expressional (mood, such as homesickness and joy). Perceiving factual meaning requires “everyday familiarity with objects and events”, while perceiving expressional meaning requires “a certain sensitivity, but this sensitivity is still part of [...] practical experience, that is, of [...] everyday familiarity with objects and events” (ibid.). The former is more objective (describes what the image is *of*) and the latter is more subjective (describes what the image is *about*) but they are both still grounded in our knowledge of the world. For example, in George Lambdin’s “Consecration” (Figure 2.6), the factual meaning contains objects such as a woman, a man and a sword and events such as a woman kissing a sword, a man holding a flower or a man looking at a woman. The expressional meaning can be a simple feeling such as ‘romantic’ or ‘gloomy’. Most humans should be able to recover the pre-iconographic meaning; recognising simple objects and moods in an image requires some elementary familiarity with culture (e.g. knowing what a man, or a sword looks like).

At the *iconographic* level, one can derive “secondary or conventional matter”, which requires familiarity with culture, beyond that of everyday objects or events. Panofsky does not distinguish between ofness and aboutness at this level, but Shatford (1986) does. The factual (‘ofness’) meaning at this level for the same picture could be ‘a Yankee library’, ‘couple during the American Civil War’, ‘a man dressed as a Union Army officer’ or ‘a woman saying goodbye to her beloved Union army officer, who is preparing to leave for the battlefield’. The expressional (‘aboutness’) meaning could be ‘the feeling of parting from a loved one for a noble cause’. People without a relevant cultural background would fail to recover the iconographic meaning.



Figure 2.6: “The Consecration” by George Lambdin (1865)

19thcentury beforeheleavesforwar black bookcase
 kiss commitment corset couple courting crinolineperiod
 genderreversal greydress greydressissymbolfor home
 man manandwoman **military** militaryofficer
 furnishings
 romantic rose **soldier** soldiers study books
 bookshelf canvas ceremony cinched **civilwar**
 forestgreen civilwarsoldier husbandandwife interior
 lady **library** librarysetting luxurious violet
 darkred **domestic** dress examine flirt present
 mutual orientalcarpet **painting** periodcostume
sword thecivilwar two uniform portrait
 womanholdingsword womaninalongdress women
 yankeeuniform weaponry woman upperclass red

The *iconological* level of interpretation involves recovering the “intrinsic meaning of content”, which requires a good understanding of the artist, the social circumstances that the image is set against, symbolism and so on. People without specific historical knowledge of the American Civil War would fail to arrive at successful iconological interpretation. This level has no ofness-aboutness distinction because relations are much too complex at this stage, so iconology is a general discussion of the painting. In the image shown in Figure 2.6, one could open an iconological discussion by mentioning that the painting was created during the last year of the Civil War, that is, one year before the South reconciled with the North. The woman, who is dressed in grey, which is the colour of the Confederation army (South), might symbolise the realisation of the South that it should be re-united with the North, symbolised by the man who is dressed in a Union army uniform. The artist preferred to give his account of the war through domestic settings, leaving the cruelty of war for journalists to describe.

The tag cloud of the image in question contains tags that reveal an understanding at different levels (e.g. *pre-iconographic tags*: “woman”, “couple”, “sword”, “kiss”, “library”, “man and woman” etc. (‘of’) and “romantic” (‘about’); *iconographic tags*: “soldier”, “civil war”, “yankee uniform” etc. (‘of’) and “flirt”, “commitment” (‘about’); *iconological tags*: “grey dress is symbol for”, “gender reversal” etc). In an analysis of tagged images, Peters (2007) also observes that, in tagging systems, tags are assigned at all three levels of image interpretation.

2.4 Summary

In this chapter, I provided theoretical background that will facilitate understanding of subsequent chapters. First, I described the image tagging corpus used in this research and then I discussed the nature of digital images as documents. Finally, I analysed the image-tag relationship, focusing on the role of tags as indicators of a document's subject, while distinguishing between theme and rheme, ofness and aboutness, as well as pre-iconography, iconography and iconology.

Chapter 3

Tags as words

This chapter describes some initial measurements made to investigate the nature of individual tags in the dataset. The probability distribution of the tagging data and the relationship between tags and natural language words will be discussed.

3.1 Probability distribution of tags

As mentioned in Section 1.2, a large enough folksonomy is expected to exhibit complex system properties. Even though folksonomy vocabulary is uncontrolled, it stabilises over time when users reach a consensus on tag assignment (Halpin et al., 2007). Bollen and Halpin (2009) report that such an agreement arises even in the absence of a tag recommendation mechanism; users tend to use the same tags with or without exposure to each other's tagging activity. In practical terms, a consensus means that the frequencies of tags over an entire folksonomy follow a power law distribution, starting with very high values for the first few tags and dropping dramatically to a 'long tail' of rarely used tags. That is, some popular tags become even more popular in a rich-get-richer fashion, thereby reflecting how users conceptualise objects in the folksonomy.

The Steve corpus of tagged images used in this research (see §2.1) strongly confirms the above expectations. The 447,532 total occurrences (tokens) of the 65,065 unique tags (types) are distributed far from equally or normally; a small set of tags account for a large portion of the probability distribution. Figure 3.1a shows the frequencies of all distinct tags in the entire dataset, sorted from the most to the least popular. The red curve is almost aligned with the y and x axes, showing the high peak ('head') and the long tail respectively. Figure 3.1b shows that, in a log-log scale, the line tends to become straight. Figures 3.1c and 3.1d are the same plots as (a) and (b), only zoomed to the 500 most popular tags across the folksonomy.

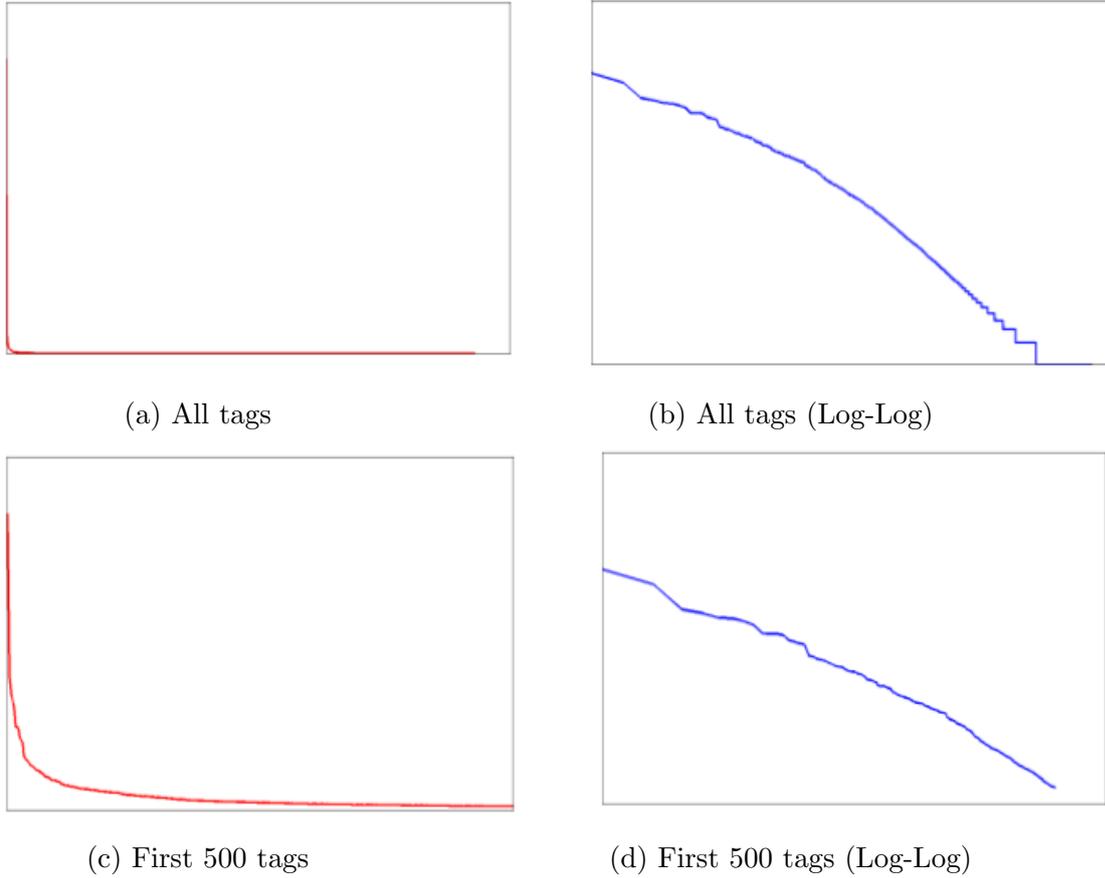


Figure 3.1: Tag distribution on Steve Tagger

Tags in the Steve folksonomy, as in many other folksonomies in the literature, follow a power-law distribution, known in linguistics as Zipf’s law (Zipf, 1932, 1949). For natural languages, Zipf’s law predicts that if words in a corpus are ranked from the most to the least frequently occurring, this frequency is a power-law function (f) of the word’s rank (k), that is $f(k) = \frac{1}{k^s}$, where s is a constant close to 1. For the Steve dataset, I estimated the s parameter with the Levenberg-Marquardt algorithm (Levenberg, 1944) for non-linear least-squares curve fitting and found that $s = 1.03$. As can be seen in Figure 3.2, the distribution of Steve tags can be described using a zipfian curve, as observed for words in text corpora (e.g. in the Brown Corpus; Kucera and Francis 1979; Francis and Kucera 1982).¹

¹Fitting a zipfian curve to a distribution of words from a corpus tends to produce error in the top and bottom ranks. The Zipf-Mandelbrot law (Mandelbrot, 1965), which requires two instead of one parameter, is typically used to provide a better fit. In this thesis, the traditional Zipf’s law was considered enough to demonstrate a tendency for tags to follow a power-law distribution.

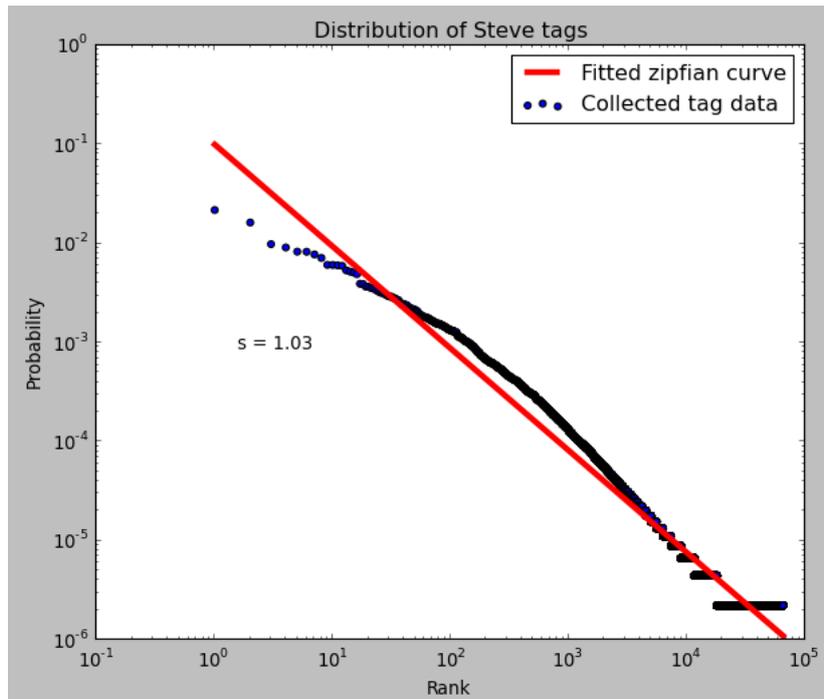


Figure 3.2: Folksonomy tags with fitted Zipf distribution

A similar way to describe the distribution of tags in the Steve corpus is by means of Heaps' law (Heaps 1978, pp. 206-208), originally formulated by Herdan (1960). According to this law, the size of a vocabulary (number of types) grows at a slower rate as text size (number of tokens) increases. For instance, by observing the first 100 word tokens of a corpus, we can discover, say, 85 distinct word types, among which we are very likely to discover some of the most frequently occurring words in the corpus. As we keep observing more tokens, the vocabulary grows but it becomes more difficult to discover new word types; a large number of the tokens encountered will belong to already known vocabulary items. The law can be formally described as follows:

$$V(n) = Kn^\beta \tag{3.1}$$

where V is the size of the vocabulary discovered in a sample of n tokens; K and β are parameters.

This law is a direct consequence of a dataset that follows a Zipfian distribution. We would therefore expect that the Steve corpus obeys Heaps' law too. To demonstrate this property, I performed a Monte Carlo simulation by treating the 447,532 tag tokens in the Steve corpus as a population and drawing samples from it in order to observe the rate of vocabulary growth. Starting with an initial discovered vocabulary of size 0, I drew samples of 500 tokens without replacement, until the population was exhausted. At each sampling, I recorded the number of new vocabulary items discovered. Each sampling was performed 10 times in order to reduce unwanted random effects, and the number recorded at the end of a sampling was the average of the numbers revealed by each one of the 10 samples. The results, shown in Figure 3.3 confirm our expectations.

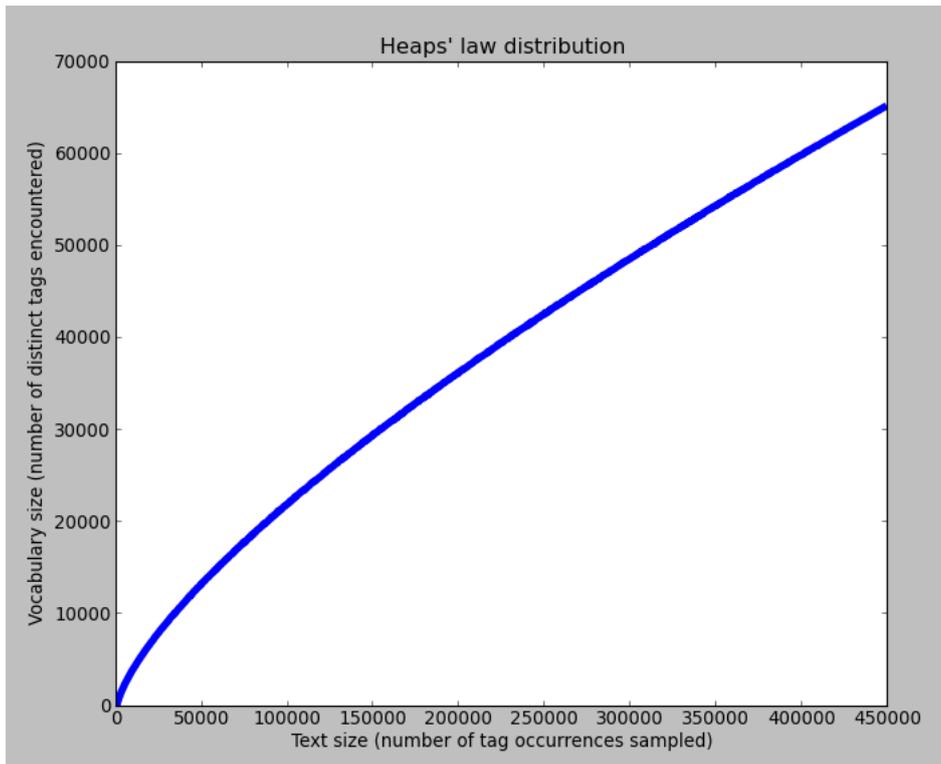


Figure 3.3: Rate of discovery of new tags

3.1.1 Head and tail of distribution

Head

As power laws predict, a small percentage of types (distinct tags) can account for a large percentage of tokens (occurrences). This is called the *head* of the distribution: if the y axis of the distribution represents probabilities, the head could be seen as the peak of the (highly skewed) probability mass graph. In the Steve dataset, the most popular tag (“black”) has been used 9,742 times, that is 2.2% of all 447,532 times that any tag was used (see Table 3.1). The second most popular tag (“white”) accounts for 1.6% of all occurrences, while the top five tags together cover 6.5% of all tokens (see Table 3.2). It takes 388 tags to account for half of the occurrences in the corpus. In the Brown Corpus of American English (Kucera and Francis, 1979), the top word (“the”) occurs almost 7% of the time, the next word (“of”) accounts for over 3.5% of the occurrences while half of the corpus is covered by the top 135 words (Francis and Kucera, 1982). The Steve dataset follows the same trend but the decline of frequencies from the head to the tail is less steep. One explanation could be that the Brown corpus has a different type/token ratio because it is at least twice as large (1,014,312 tokens) as Steve. Another explanation could be that the most popular Steve tags are not closed-class words (e.g. articles, prepositions, pronouns etc.), which one would expect to see in the top ranks of a natural language zipfian distribution. Zipf (1949) suggested that informal text exhibits a steeper decline in the head than in the rest of the distribution because of the wider use of personal pronouns, which increases the number of closed-class words in the top ranks.

Tag occurrences in the head of the Steve corpus have a steeper slope than that predicted by the Pareto principle, an empirical law used in the social sciences to describe

many phenomena that follow a power law distribution (e.g. incomes across individuals). According to the Pareto principle, also referred to as the 80-20 rule, 80% of events caused (here tag occurrences) can be traced to 20% of the causes (here tag types). In the Steve corpus, 80% of tokens are accounted for by only 7.3% of the distinct tags, while the top 20% of tag types account for 87% of tag occurrences (Table 3.3).

Top ranks	Tag tokens	
	%	#
1 ('black')	2.2	9,742
2 ('white')	1.6	7,289
3 ('tree')	1	4,412
4 ('brown')	0.9	4,084
5 ('man')	0.8	3,708
TOTAL	100	447,532

Table 3.1: Percentage (%) and number (#) of tokens in the top five ranks

Top ranks	Tag tokens		Tag types	
	%	#	%	#
1 - 5	6.5	29,235	0.01	5
1 - 10	10.1	45,059	0.02	10
1 - 50	23.0	102,923	0.08	50
1 - 100	31.1	139,083	0.15	100
1 - 200	40.6	181,591	0.31	200
1 - 388	50.0	223,683	0.6	388
1 - 500	52.7	240,258	0.77	500
1 - 1,000	63.3	283,432	1.54	1,000
1 - 2,000	71.7	320,722	3.07	2,000
1 - 4,750	80.0	358,143	7.3	4,750
1 - 10,000	85.6	383,046	15.37	10,000
1 - 13,010	87.3	390,627	20.0	13,010
TOTAL	100	447,532	100	65,065

Table 3.2: Cumulative percentage (%) and number (#) of tokens and types in the top ranks

Bottom ranks	Tag tokens		Tag types	
	%	#	%	#
6,257 - 7,320 (count 5)	1.2	5,320	1.6	1,064
7,321 - 8,862 (count 4)	1.4	6,168	2.4	1,542
8,863 - 11,561 (count 3)	1.8	8,097	4.1	2,699
11,562 - 17,860 (count 2)	2.8	12,598	9.7	6,299
17,861 - 65,065 (hapaxes)	10.5	47,205	72.6	47,205
6,257 - 65,065 (counts 1-5)	17.7	79,388	90.4	58,809
TOTAL	100	447,532	100	65,065

Table 3.3: Cumulative percentage (%) and number (#) of tokens and types in the bottom ranks

Tail

The bottom ranks, also known as the *tail*, of a power law distribution exhibit the opposite phenomenon from that of the head: a large percentage of low-frequency types can account for only a small percentage of total occurrences. For instance, hapax legomena (i.e. tags that occur only once in the entire corpus) are 72.6% of the distinct tags in the Steve corpus but they account for only 10.5% of tokens in the dataset. The bottom 90.4% of tag types in this corpus (i.e. those that occur at most five times) account for only 17.7% of occurrences (Table 3.3). The Steve corpus has a heavier tail than the Brown Corpus, where approximately 50% of the vocabulary items are hapax legomena.

To provide an explanation for the heavier-than-expected tail in the Steve data, I examined the nature of the hapax legomena. After initial manual inspection, it was obvious that a large number of hapaxes were tags such as “japantreestemple”, “hornwithcase” and “capitallionsgothicspain”, which consist of two or more words (i.e. “japan trees temple”, “horn with case” and “capital lions gothic spain” respectively). Multi-word tags in this dataset lack word boundary markers, such as underscores (“horn_with_case”), camel-case (“hornWithCase”) or spaces (“horn with case”), because of whitespace elimination that took place on the Steve Tagger platform after a user had submitted a tag (e.g. “symbolisms in tapestry” is converted to “symbolismsintapestry”, with the latter form appearing in a given tag cloud).

Multi-word tags may result from users’ need to express a concept in more than one word or even from their misunderstanding of the tagging interface (e.g. tag separators) leading them to submit one multi-word tag (e.g. “chair night shirt france”) where one tag per word (“chair”, “night”, “shirt”, “france”) was intended. If multi-word tags account for a large percentage of hapax legomena, this may explain why the tail of the tag distribution is heavier than that observed in text tokenised into individual words. In order to quantify this informal observation, I created a system that identifies and normalises multi-word tags, adding word boundaries. The process was as follows:

1. **creating unseen subcorpus:** Approximately 1/10 of Steve tags were randomly selected, separated from the corpus and kept aside for testing.
2. **choosing a lexicon:** Learning word boundaries necessitates a notion of what a ‘word’ is. In this research, any unlemmatised vocabulary item found in a text corpus was considered a word. The ideal corpus for this task was Wikipedia, because it

one of the largest available corpora for English and covers a wide variety of topics. I used Wikipedia via the Wikiwoods corpus (release 1010; Flickinger et al. 2010)². Although the corpus is known for its HPSG³ syntactico-semantic annotations, it is also distributed as simple text which has been cleaned from Wikipedia annotations and segmented into sentences. As expected, Wikiwoods has a large vocabulary (5,051,015 word types).

3. **constructing a language model:** Unigrams, bigrams and trigrams were extracted from Wikiwoods to be used for determining the best segmentations (e.g. “insectface-dragonchina” can be segmented as “insect face dragon china”, “in sect face dragon china”, “insect face drag on china” etc.)
4. **learning word boundaries:** Every tag that was not found in the Wikiwoods lexicon was a candidate for splitting. Particular tokens from Wikiwoods were treated as stopwords on the basis of heuristics (e.g. having 2 characters and at the same time occurring less than 3,000 times in the corpus). The heuristics were set empirically, through observation of the training data. The different possible segmentations of a candidate word (e.g. “sculpture stone”, “sculptures tone” and “sculpture st one” for the tag “sculpturestone”) were ordered according to probability and the most probable segmentation was kept. Unigram probabilities of the component words were found to be better predictors of the quality of a segmentation than higher ngrams. This can be justified on the basis of the observation that a large portion of multi-word tags were composed of words that do not necessarily occur in the same order in corpora. The list of words that form a multi-word tag do not necessarily have syntactic relations holding between them; they are often just a list of semantically related but not syntactically constrained words (e.g. “deityhinduhinduism” analysed as “deity hindu hinduism”).
5. **evaluating the system:** To evaluate the quality of multi-word tag segmentation, I hand-picked 100 cases of tags from the test corpus (step 1 above) that required splitting and manually produced the canonical form. The system’s most likely segmentation agreed with the manual segmentation of each multi-word tag in 98% of the cases. The results were considered adequate for this task, so no further improvement was attempted.

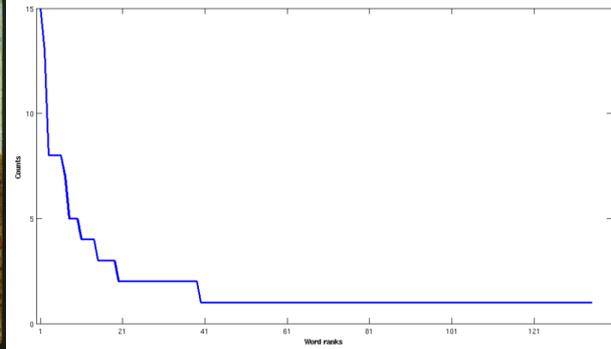
Based on the above system, almost half (35,520) of the hapax legomena in the Steve corpus are multi-word tags. Almost 99% of multi-word tag types occur less than 10 times in the corpus, with 85.6% being hapax legomena. The number of multi-word tags in the low ranks of the distribution is, then, enough to explain the heavy tail observed.

3.1.2 Resource-specific tag distribution

A power law distribution is not an exclusive property of the entire tagging corpus; it also tends to hold between tags associated with a particular resource, provided that the tag cloud is large enough. An example can be seen in Figure 3.4.

²<http://ltr.uio.no/wikiwoods/1010/>

³Head-driven Phrase Structure Grammar



19thcentury afternoon asymmetrical asymmetricalwithvividcontent aurinko **beach** beachboyswithsailboats blue bluesky
 bluewater **boat** boatbuilders boatbuilding boatmaking **boats** boy **boys** boysinhats brothers bygone
 calm capeann chikdren child **children** coast constructingsail craftyboys crap creating europeancoast
 fatherandson gloucester gull hat **hats** havet **homer** horizon inlet jetties kesapaiva koast lamer light lightgray
 lighthearted manyvaluesofbrownused massachusetts model **models** mothering mothers muted naturalistic **newengland**
 newenglandboyswithsailboat niceday **ocean** oilonpanel oilpainting orange **painting** parents
 parentsandchildren peaceful pebbles purjeveneranta realism **red** redshirt relaxation rock
rocks rocky rockybeach rockyshoreline sailbaat **sailboat** **sailboats** sailing sailingship
 sailingships sails sand scene **schooner** schooners **sea** seacoast seagull seagulls **seascape** seashore
 seaside seasidelandscape **seaweed** see serene sharing ship shipmodels ships shirt shore sienna silver **sky** solntze
 soothing strand **strawhats** summer summeratsea summery sun sunny **sunshine** tan teal
 thispaintingisveryaestheticallypleasingtomeigetafeelingofpeacefromjustlookingatthemanandboy toy toyboat **toyboats** toys
 viewoftheocean vividblues water watershowsatmosphericperspective wave white wind winslowhomer worthingtonwittridge

Figure 3.4: Distribution of tags on painting “The Boat Builders” by Winslow Homer (1873); Graph of all tags used on resource (x axis) plotted against the number of times they have been used (by every user on this resource)

3.2 Tag occurrences in natural language text

As mentioned in Section 1.1, tags are uncontrolled vocabulary; they are not selected from a pre-approved list of index terms, as in the case of professional cataloguing, but are simply any sequence of characters that could help render the resource searchable. These characters can be letters, numbers, punctuation, mathematical notation or any other symbol supported by the tagging system. When a user creates a tag, they combine characters in order to communicate an aspect of the resource to themselves or others for the purposes of future retrieval. We would therefore expect that a tag is equivalent to a word, that is a known entry in a language’s lexicon, whose meaning can capture the aspect of the resource that the user intends to communicate. However, nonsense strings (e.g. ‘bsrtooo’) or strings that have an obscure meaning (e.g. ‘3RST_z(1)’) can still be valid tags and facilitate one’s retrieval of a document. In addition, it is possible that the nature of the tagging task (e.g. its speed) might promote the emergence of novel forms (e.g. emoticons such as ‘:’) or ‘:D’) or spelling variants (e.g. ‘img’ for ‘image’,

‘gr8’ for ‘great’ etc.), as has been observed in internet slang (Crystal, 2001) and text messaging (Crystal, 2009). Hence, to further examine whether tags behave like natural language words, it is important to quantify the extent to which folksonomy tags are strings classifiable as words.

As mentioned in Section 3.1.1, in these experiments, a tag is considered equivalent to a word if it is found as an unlemmatised vocabulary type in one of the two text corpora used, Wikiwoods (Flickinger et al., 2010) and the British National Corpus (BNC) (Aston and Burnard, 1998), which comprise 5,051,015 and 939,028 types respectively. Among the 65,065 tag types in Steve, 22,024 (33.8%) were word types in Wikiwoods. However, among the rest of the tag types (i.e. 43,041), 96.4% (41,473) were found to be multi-word tags, whose component words appeared in Wikiwoods. This means that 63,497 tag types (97.6% of all) were either words or combinations of words found in the text corpus.⁴ BNC is a much smaller corpus, so only 26.9% of tags were identical to words, but this rises to 90.6% if multi-word tags are also considered.

3.3 Tags and parts of speech

The vast majority of tags in folksonomies have been found to be nouns and adjectives: Guy and Tonkin (2006) revealed that 90% of the tags submitted on Delicious and Flickr are nouns. Spiteri (2007) found that nouns account for 88% of tags on Delicious tags, 71% on bookmarking website Furl⁵, and 86% on advertising platform Technorati.⁶ Adjectives in the three systems were found to cover 6%, 6% and 3% of the tags respectively. The author’s explanation for the dominance of nouns was that the tags people use in order to annotate resources tend to represent things. In particular, tags tend to indicate things that the image is *of* or *about* (see §2.3).

However, deciding on the part-of-speech (POS) tag of a word in isolation is very problematic. Words are POS-tagged, either manually or automatically, with respect to their neighbouring words in text. For example, the word “green”, can function as an adjective in the sentence “I like green vegetables” and as a noun in the sentence “Green is my favourite colour”. To overcome this limitation, I used a list of unlemmatised BNC words (Kilgarriff, 1995)⁷, each one of which is associated with a frequency-weighted list of POS tags that it has been labelled with in the text corpus by the automatic tagger estimate. For example, the word “olive” has functioned as an adjective (AJ0) 422 of the times it has appeared in text, as adjective or singular common noun (AJ0-NN1) 285 times, as proper noun (NP0) 142 times, as singular common noun (NN1) 41 times and as singular common noun or proper noun (NN1-NP0) 34 times.⁸

In order to measure the percentage of nouns and adjectives in the Steve dataset, I defined three conditions for each one of the two categories: ‘usually’, ‘possibly’ and

⁴A sample of the remaining 2.4% suggested that the tags which were neither words nor concatenations of words were nonsensical (e.g. “hjhjk”, “zxczxcfz”, “blahhhhhh”, “bzzzzzzzzzz”), misspellings (e.g. “artneauveau”, “eating”, “worhsipping”, “whelthrown”, “himduism”, “grpes”), non-English words (e.g. “zhangkunyi”, “calligrafie”) sometimes written in non-Ascii characters, or tags with obscure meaning (e.g. “1986326ab”).

⁵<http://web.archive.org/web/20081231221904/http://www.furl.net/>

⁶<http://technorati.com/>

⁷<http://www.kilgarriff.co.uk/bnc-readme.html>

⁸The POS tagset used is CLAWS for BNC, available on <http://www.kilgarriff.co.uk/BNClists/poscodes.html>

	noun	adjective	Noun or adjective
<i>usually</i>	11,919 (68.2%)	3,022 (17.3%)	14,941 (85.5%)
<i>possibly</i>	13,589 (77.8%)	5,131 (29.4%)	16,996 (97.3%)
<i>never</i>	3,881 (22.2%)	12,339 (70.6%)	474 (2.7%)

Table 3.4: POS tags in the Steve corpus

‘never’. A word is *usually* a noun if its most likely POS tag in text derives from the noun list ‘NN0’, ‘NN1’, ‘NN2’, ‘NN0’, ‘NN1-NP0’, ‘NN1-VVB’, ‘NN1-VVG’, ‘NN2-VVZ’, and usually an adjective if its most likely POS tag comes from the adjective list ‘AJ0’, ‘AJC’, ‘AJS’, ‘AJ0-AV0’, ‘AJ0-NN1’, ‘AJ0-VVD’, ‘AJ0-VVG’, ‘AJ0-VVN’. A word is *possibly* a noun if at least one of the POS tags it has been associated with, regardless of frequency, is in the above noun list, and possibly an adjective if at least one of its POS tags is in the above adjective list. Finally, a word in *never* a noun if none of its tags are in the noun list and never an adjective if none of its tags are in the adjective list. As can be seen in Table 3.4, among the 17,470 tags in the Steve corpus that occur as single words in BNC, 68.2% are usually nouns and 17.3% are usually adjectives. 77.8% of the Steve tags that are BNC words are *possibly* nouns and 29.4% are possibly adjectives. A large proportion of BNC-encountered Steve tags (85.5%) are usually found as a noun or an adjective, while the vast majority (97.3%) can be found as either of these two POS tags. The remaining 2.7% (474), which are never nouns or adjectives, tend to belong to the following parts of speech:

- **verbs:** e.g. ‘eat’, ‘know’, ‘made’
- **prepositions:** e.g. ‘of’, ‘with’
- **adverbs:** e.g. ‘together’, ‘quite’
- **dates:** e.g. ‘1921’, ‘2003’
- **unclassified:** e.g. ‘rhomb’, ‘achoo’, ‘no4’

3.4 Tags and word categories

To gain a better understanding of the concepts that tags are meant to annotate, it is interesting to explore what semantic categories image tags are associated with. Mathes (2004) reports that the top 150 tags on Flickr label “common subjects of photos” such as animals, friends, cities, gardens and so on. Approximately 25% of Flickr tags are proper names for places, while colours and years are also popular. The author also found that some tags were subjective and personal (e.g. “cute”, “me”), which shows “the importance of individuality and ego for these systems to work”. In the Steve dataset, I intuitively categorised the top 200 tags into 11 semantic fields. As seen in Table 3.5, tags are words classifiable under: artistic techniques & materials (21.5%), colours (19%), objects (16.5%), nature & landscapes (19.5%), people (9.5%), countries & nationalities (4.5%), subjective terms (3.5%), descriptive terms (3.5%), events (1.5%), abstract terms (0.5%) and unclassified tags (0.5%). These word categories are more varied than those of Flickr tags. Based on the different levels of interpretation in art and the potentially nested

media (see Sections 2.2 and 2.3), it is unsurprising that Steve contains tags from a larger variety of semantic fields. The semantic fields covered by the most-frequently occurring Steve tags are common topics in art discourse, which provides some re-assurance that comparing the function of tags in a tag cloud with the function of words in text is not unreasonable.

Table 3.5: Word categories in the top 200 Steve tags

artistic techniques & materials	silver, wood, china, paper, cloth, metal, ceramic, ink, fabric, carved, pottery, watercolor, abstract, line, design, sketch, sculpture, portrait, exhibition, painting, pattern, print, circle, color, square, figure, floral, geometric, rectangle, shape, text, architecture, writing, dot, arch, antique, decorative, curve, round, colorful, bronze, surface, ivory
nature & landscapes	tree, water, mountain, landscape, sky, cloud, river, flower, grass, leaf, trees, horse, stone, hill, nature, rock, forest, bird, plant, sea, clouds, lavender, branch, shadow, light, lake, natural, dog, sand, ocean, flowers, bush, field, leaves, shore, wave, town, city, village
colours	black, white, brown, tan, red, blue, beige, sienna, green, yellow, gold, gray, orange, lightgray, royalblue, maroon, slategray, pink, orangered, darkred, grey, olive, darkorange, dark, lightblue, lightpink, darkslategray, darkgreen, forestgreen, darkgray, seagreen, purple, darkblue, turquoise, teal, slateblue, violet, navy
objects	building, house, dress, boat, hat, bridge, pencil, table, window, statue, road, book, wall, windows, chair, door, bowl, coat, vase, stick, ship, castle, church, fence, home, tower, cap, path, gown, cross, roof, charcoal, sword
people	man, woman, face, child, lady, eyes, hair, men, nose, nude, female, girl, women, head, beard, male, hand, hands, boy
countries & nationalities	france, unitedstates, korea, french, india, chinese, japanese, american, asian
subjective tags	nice, good, beautiful, normal, best, wonderful, super
descriptive tags	old, ancient, simple, modern, religious, oriental, happy
events	sitting, standing, surprise
abstract tags	fashion
unclassified tags	other

3.5 Summary

In this chapter, I explored the nature of individual tags in the Steve folksonomy. I showed that tags follow a power-law distribution and are almost exclusively words (or concatenations of words) found in corpora. I also demonstrated that the vast majority of

tags act as nouns or adjectives and that the most frequent tags in the folksonomy cover subjects common in art discourse.

Chapter 4

Tags in combinations

In this chapter, I examine whether it is possible to draw parallels between the way tags combine to form tag clouds and the way words combine to form larger units in text. In Section 4.1, I provide some theoretical background on relationships between words in natural language, which I will later apply to the investigation of combinatory patterns in tag clouds. In Sections 4.2 and 4.3, I describe the distributional properties (co-occurrence and similarity respectively) of tags in tag clouds and compare them with those of words in text.

4.1 Relationships between words

4.1.1 Syntagmatic and paradigmatic relations

Structural linguists, following Saussure's *Cours de linguistique générale* (1916), have seen meaning as internal to the language (unlike theories based on intension, extension or usage); a word's *valeur* (Saussure, 1916), that is 'meaning' or 'function' in the language, is defined by means of the word's relation to other words. Saussure distinguished between two types of relations, *syntagmatic*, and associative, widely referred to as *paradigmatic*.¹ Syntagmatic relations are combinatorial, also known as 'horizontal'. For instance, the words in sentence (1) below relate to each other by combination. The specific nature of these relations is not specified; they could be phonological, (morpho-)syntactic, semantic, or simple co-occurrences. *Paradigmatic* relations, on the other hand, are substitutional, also known as 'vertical'. They are relations between words that can substitute for one another in particular linguistic contexts. For instance, we could plausibly construct sentence (2) below but not sentences (3) or (4); "cat" and "dog" are paradigmatically related but "cat" and "interesting" are not.

- (1) A fast cat was chasing another animal.
- (2) A fast dog was chasing another animal
- (3) *A fast interesting was chasing another animal.
- (4) ?A fast idea was chasing another animal.

¹The term 'paradigmatic' was introduced by structural linguist Roman Jakobson (1941).

Saussure suggested that syntagmatic relations occur *in praesentia*, since they are inferred by examining how words are arranged when simultaneously present in particular linguistic contexts ('syntagms'). Paradigmatic relations occur *in absentia* because words that could replace one another in syntagms do not need to be present at the same time; a paradigmatic set, that is a set of inter-substitutable words, simply contains possibilities given syntagmatic restrictions. It is obvious that paradigmatic relations cannot be observed in what Saussure called 'parole' (i.e. specific instances of language use) but they are part of *langue* (i.e. knowledge of language). What is less obvious, though, is that syntagmatic relations also belong to *langue*, so they cannot be inferred from any individual syntagm (e.g. utterance, written sentence etc.) but through examination of a large collection of syntagms: a human can learn syntagmatic relations (combinatory restrictions) by being exposed to various instances of *parole*. By analogy, a machine can learn syntagmatic relations by examining a text corpus containing many instances of parole: "In an individual text, neither repeated syntagmatic relations, nor any paradigmatic relations at all, are observable" (Stubbs, 2008); what corpus linguists call a 'concordance' (i.e. a collection of immediate linguistic contexts for a particular word) "makes visible repeated events" (Stubbs 2008; emphasis mine).

Distributional semantics

Syntagmatic and paradigmatic relations between words have been discussed in the context of the *distributional hypothesis*, whereby a word's meaning can be described by the linguistic contexts in which the word occurs (or does not occur); as Firth (1957) put it: "You shall know a word by the company it keeps". Each word can be represented by a vector whose elements are numbers that indicate how often the word is found in a particular context (e.g. in a given grammatical construction, in a sentence which contains a particular word and so on). This vector can be seen as the word's 'distributional meaning': "Distributional models are models of word meaning. Not the meanings that are in our heads, and not the meanings that are out there in the world, but the meanings that are in the text." (Sahlgren, 2008). Another way to see distributional meaning is with Zellig Harris' observation that similar words occur in similar (linguistic) contexts (Harris, 1954). For example, all words that are plausible objects of the verb 'cook', tend to belong to the same semantic category (i.e. that of foods). In practical terms, we can obtain semantic similarity estimates between two words by comparing their feature vectors.

Sahlgren (2008) argues that feature vectors in distributional semantics reveal syntagmatic relations. For instance, after constructing a vector representation for the word "cat" based on its neighbouring words in sentences, we can obtain an estimate of the combinatory constraints, that is syntagmatic relations, in which "cat" participates. If the vector for "cat" is compared to all other vectors (e.g. those of "dog", "sofa", "flower" etc.) obtained from text, then a similarity vector can be created for the word "cat", with features being other words and elements being similarity scores. Sahlgren (2008) suggests that similarity vectors reveal paradigmatic relations. The words found to be the most similar to "cat" can be said to belong to the same paradigmatic set. Hence, using a corpus, one can construct both *syntagmatic vectors* and *paradigmatic vectors*, the latter deriving from the former.

It is important to clarify that syntagmatic vectors can be constructed with features of any kind. For example, the vector representation of the word "cat" can be learnt by asking questions as varied as: 'How often does "cat" co-occur with each other word?',

‘How often is “cat” the subject of each verb?’, ‘How often is it preceded by each adjective that indicates motion?’ and so on. If the second question is used as a feature, then the syntagmatic relations obtained will be combinatorial preferences for a word *with respect to* what verbs “cat” is often the subject of. Likewise, similarity (paradigmatic) vectors derived from the above syntagmatic vectors will show which words can substitute for “cat” *with respect to* the verbs they are often the subject of.

In the context of folksonomy, Cattuto et al. (2008) explain how syntagmatic and paradigmatic relations can be extracted from tags; the former can be learnt from co-occurrence patterns of pairs of tags that annotate individual resources (i.e. which tags co-occur in tag clouds), while the latter can be learnt by comparing co-occurrence vectors for similarity. As the authors explain, a co-occurrence, or syntagmatic, vector for a given tag reveals relationships that constrain the way a tag combines with others in resources, while a similarity, or paradigmatic, vector reveals the relationships that make the tag replaceable by others in resources.

4.1.2 Lexical cohesion

Relationships between words have also been discussed by linguists within the theory of lexical cohesion, a property of discourse whereby content words across sentences can work in tandem to hold text together as one coherent unit. Before I explain how lexical cohesion applies to the investigation of tag combinations in folksonomy, I will provide a brief overview of two related concepts, *coherence* and *cohesion*.

Coherence, also referred to as ‘texture’, is “the organization of discourse with all elements present and fitting together logically” (Hinkel, 2004), or a “continuity of senses” (de Beaugrande and Dressler, 1981). More formally, it is a “property of discourse formed through the interpretation of each individual sentence relative to the interpretation of other sentences” (van Dijk 1980: 93). To perceive a piece of discourse as coherent, one needs to study particular features of the text, known as *text-based* (i.e. linguistic) features, or to possess the appropriate background knowledge (*reader-based* features) (Johns 1986: 247).

Text-based features that explain coherence have been described within the theory of cohesion. First introduced by Halliday and Hasan (1976), cohesion can be defined as “a set of lexicogrammatical systems that have evolved specifically as a resource for making it possible to transcend the boundaries of the clause” (Halliday 2014: p.603). Without cohesion, sentences would tend to lack coherence, thus remaining a group of unrelated sentences (Halliday and Hasan, 1976; Hinkel, 2004). Halliday and Hasan identified four categories of cohesion: **i)** reference (e.g. *Alice* likes rowing. *She* is crazy about it.), **ii)** ellipsis and substitution (e.g. They asked me which *bike* I preferred. I bought the blue *one*.), **iii)** conjunction (George is much older than the other employees. *But* the company needed someone with exceptional experience.) and **iv)** lexical cohesion (e.g. Jersey *fabric* drapes nicely. I bought it in the *textile* shop at a discount. *Sewing* doesn’t have to be expensive). Lexical cohesion might explain why a group of tags such as “phone”, “communication” and “speaking” seem to be telling a story more plausibly than a group like “dog”, “galaxy” and “microphone”. In other words, lexical cohesion might be an indication of coherence.

Cohesion and coherence in topic modelling

Cohesion, and hence coherence, of a tag cloud can be quantified using variants of techniques proposed for the evaluation of topic models. Topic models are “algorithms for discovering the main themes [topics] that pervade a large and otherwise unstructured collection of documents” (Blei et al., 2010). A standard method for discovering topics in documents is Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Steyvers and Griffiths, 2007), a generative probabilistic model whereby documents are seen as multinomial distributions of latent topics and topics are seen as multinomial distributions of words. For a pre-determined number of topics, LDA can estimate the proportion of different topics in each document and the proportion of different words in each topic. A learnt topic is a distribution of words, much like a tag cloud, thus, methods used for evaluating the quality of the former might provide a basis for the evaluation of the latter.

Newman et al. (2010) argue that the quality of a topic can be determined by its coherence. According to the authors, coherence is equivalent to the topic’s “semantic interpretability”, that is the ability of a human to describe the topic with a short label. To measure the coherence of a set of topics, they conducted an experiment with nine participants who were asked to indicate how “useful” (coherent) each topic was on a 3-point scale. The participants reached a high inter-annotator agreement², which legitimises the use of these judgements as standards for later evaluations. The ultimate goal of Newman et al. was to develop automatic evaluation methods which could return scores for each topic that are highly correlated with the human judgements obtained. The authors compared a variety of evaluation techniques for estimating topic coherence scores based on three different resources: Google³, Wordnet (Fellbaum, 1998) and Wikipedia⁴. All topics were restricted to the 10 words with the highest probability in the topic. Using Google search, they submitted each topic as a single query and scored it based on either “titles match” (i.e. the number of distinct words in the topic that are found among the titles of the top 100 Google results) or “log hits match” (i.e. the base 10 log of the number of results returned). For Wordnet and Wikipedia, the coherence score of a topic was either the mean or the median of the “relatedness” scores D produced for each one of the 45 (i.e. $\binom{10}{2}$) possible pairs of the top 10 words in the topic. More formally:

$$\text{Mean-D-Score}(w) = \text{mean}\{D(w_i, w_j), i, j \in 1 \dots 10, i < j\} \quad (4.1)$$

$$\text{Median-D-Score}(w) = \text{median}\{D(w_i, w_j), i, j \in 1 \dots 10, i < j\} \quad (4.2)$$

In the case of Wordnet, the relatedness score was calculated in various ways, after mapping words to synsets (senses). The authors used path-based techniques (e.g. distance between two synsets in the Wordnet is-a hierarchy), information content of synsets in the hierarchy, mapping synsets to dictionary definitions and creating a vector representation for each synset. In the case of Wikipedia, words were mapped to articles and relatedness was determined using the information content of the articles (based on the number of links

²Agreement was defined as the average Spearman ρ correlation between each participant and the mean of all other participants (0.73 for topics learnt from a collection of news articles and 0.78 for topics learnt from a dataset of books). Another measurement was performed using the average Spearman ρ correlation between each participant and the median of all other participants (0.79 for news and 0.82 for books).

³<https://www.google.com>

⁴<https://en.wikipedia.org>

from one article to the other), the number of out-links shared by two articles, cosine-based document relatedness (where each article is represented as a vector of words) and the Pointwise Mutual Information (PMI) score $PMI(w_i, w_j) = \log p(w_i, w_j) - \log p(w_i)p(w_j)$ obtained by co-occurrences of w_i and w_j for each pair of topic words in a sliding 10-word window of Wikipedia text. Newman et al. found that PMI is the intrinsic evaluation measure for topic coherence that correlates the best with human judgements.

Aletras and Stevenson (2013) computed the relatedness between a pair of topic words through distributional semantics. They constructed a feature vector for each topic word using Wikipedia. The features were PMI of a given word with other words with which it co-occurs in a 5-word text window. Relatedness between two words in a topic was defined as equal to the similarity of the two corresponding vectors⁵, while coherence of the entire topic was the mean of all the pairwise word similarities (i.e. $\binom{n}{2}$, where n is the number of distinct words in a topic). Aletras and Stevenson found that the topic coherence computed with the above measure correlated with human judgements better than the measure proposed by Newman et al. (2010).

The above results show that the coherence of a topic can be reliably estimated through quantification of the relatedness between its component words. Relatedness can be calculated by means of the statistical relationships between pairs of words as manifest in text corpora. If topics are seen as analogues to tag clouds and relatedness between component words (or tags) is seen as a measure of lexical cohesion, it can be said that tag cloud coherence can be predicted through the statistical properties of its tags. In the next two sections, I use syntagmatic and paradigmatic distributional properties of tags to decide whether tag clouds are cohesive entities, capable of telling a coherent story with respect to the images they annotate.

4.2 Tag co-occurrence patterns

4.2.1 Co-occurrence vectors

To learn syntagmatic relations between tags as they appear in tag clouds of the Steve dataset, I constructed feature vectors based, initially, on simple tag co-occurrence. The process can be described as follows:

1. **setting a threshold:** discarding all the tags that appear less than 20 times in the corpus. After this step, tag types were reduced from 65,065 to 2,382, which is expected given the power-law distribution observed (see §3.1). The threshold was set not only for reasons of computational efficiency but also because very infrequent tags might not provide reliable context for distributional techniques.
2. **creating a matrix:** initialising a 2,382 by 2,382 empty matrix. Each cell $i - j$ (and its mirror cell $j - i$) was filled with the number of images that had been annotated by both the i^{th} and the j^{th} most popular tag in the folksonomy. For example, cell $0 - 1$ (and its mirror cell $1 - 0$) contained number 4,705 because the most popular tag (“black”) and the second most popular tag (“white”) co-occurred in 4,705 tag clouds.

⁵The researchers measured similarity in different ways: the cosine of the angle of the two vectors in the semantic space, Jaccard coefficient and Dice coefficient.

3. **visualising the matrix:** plotting the data on a checkerboard plot to identify possible patterns. The full graph is not presented here because of space limitations, however, focusing on the first few hundred tags (i.e. if we visualise the top-left corner of the full matrix; Figure 4.1), we can already see some clear tendencies.

In Figure 4.1 we can see two versions of the matrix, **i)** zoomed to the first 500 and **ii)** zoomed to the first 200 elements. First, we can observe that there is a diagonal, whose elements were set to zero, as the co-occurrence of a tag with itself (i.e. number of resources it appears in) was irrelevant for the task. The matrix is symmetrical with respect to this diagonal, so it could also be visualised as a triangle. More importantly, it is obvious that the top-left corner has higher values (in yellow) than the rest of the matrix, which means that simple frequency as a way to quantify tag relatedness tends to favour frequent tags.⁶

A notable exception to the bias caused by frequent terms in the matrix seems to be a vector illustrated by a blue line (both vertically and horizontally) that clearly contrasts the yellow, high-count corner of the matrix. This is the vector of the tag “exhibition”, which, despite being the 44th most popular tag in the folksonomy, does not tend to occur with other popular tags, hence resisting the bias towards high co-occurrence frequencies between popular tags. This tag is particularly dissociated with the tags in the top ranks (i.e. it co-occurs with them less often than one would expect by chance) but, at the same time, particularly associated with tags in lower ranks: for example, it co-occurs 198 times with the tag “art” (rank 293) and 146 times with tag tag “painting” (rank 64) but only 7 times with the tag “black” (rank 1), four times with the tag “white” (rank 2) and never with the tag “tree” (rank 3). High co-occurrences of “exhibition” with other tags can be seen by the yellow dots that break the blue line (either vertically or horizontally since the matrix is symmetrical).

To minimise the bias inflating co-occurrences between frequent words, it was necessary to use a measure of the *above-chance* co-occurrence of two tags, which penalises popular tags and favours high co-occurrences in less frequent tags. A measure that achieves this and has been widely used in distributional semantics is Pointwise Mutual Information (PMI), an information theoretic association measure that quantifies the statistical independence of two words co-occurring in a given context. PMI is often used in collocation extraction, to identify word combinations that have an idiosyncratic, as opposed to random, distribution. In this case, PMI can measure the dependence of two tags t_i and t_j by comparing the probability $P(t_i, t_j)$ of them occurring together in the folksonomy with the co-occurrence probability $P(t_i)P(t_j)$ that one could expect to see if they were independent. In order to calculate PMI values for tag pairs, I followed Pantel and Ravichandran’s (2004) equation, explained in steps below:

$$N = \sum_{i=1}^n \sum_{j=1}^m c_{ij} \quad (4.3)$$

For n words (here tags) and m features (here also tags), N is the corpus size, defined as the total count of all words in all contexts [i.e. number that arises from summing the square matrix horizontally (for every word) and then vertically (for every feature)]

⁶Indeed, the top five tags in the folksonomy are “black”, “white”, “tree”, “brown” and “tan”, while the top five pairs are “black”-“white”, “black”-“brown”, “black”-“tan”, “black”-“red” and “tree” “black”.

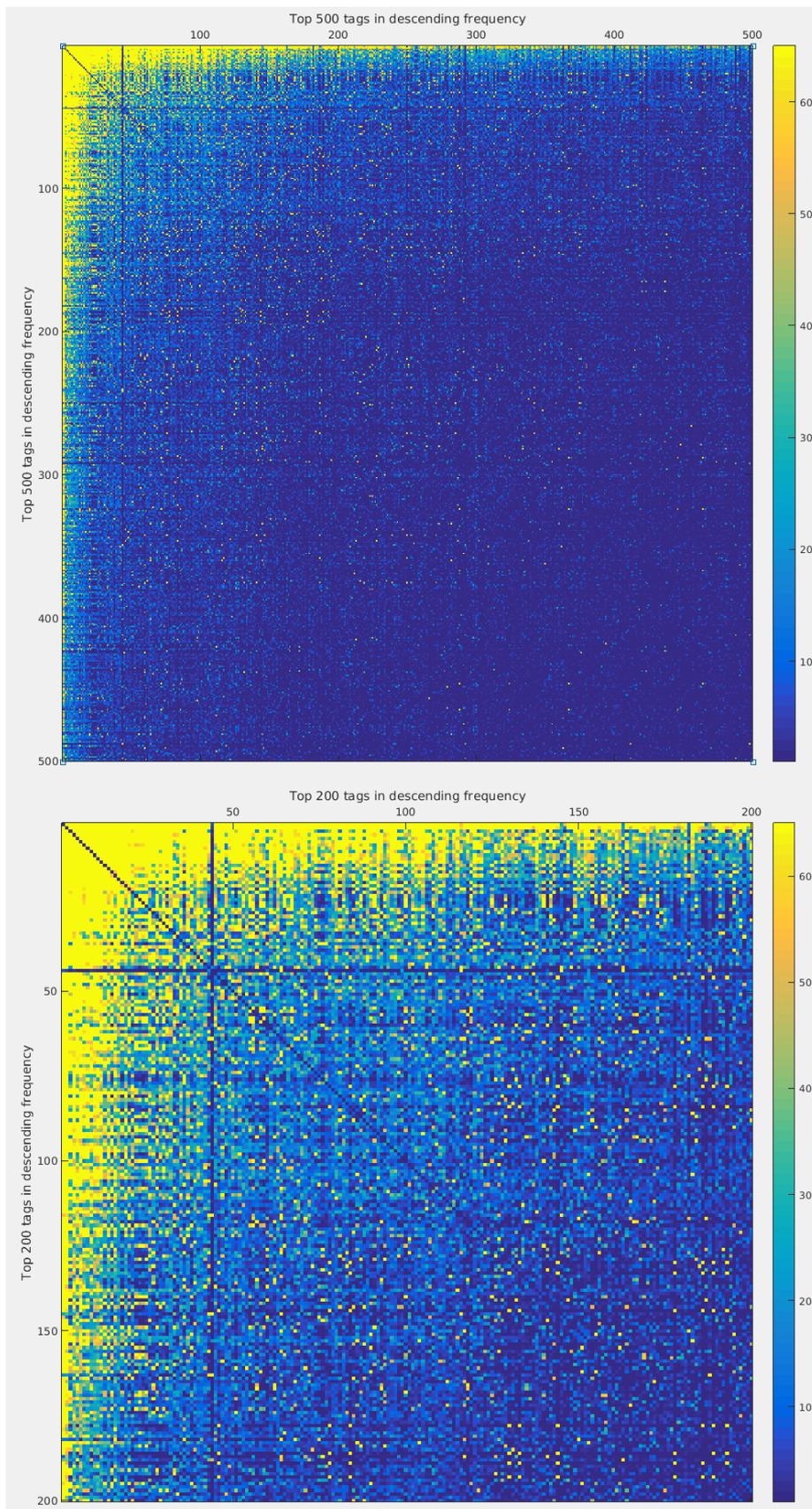


Figure 4.1: Tag co-occurrence matrices

$$F = \sum_{i=1}^n c_{if} \quad (4.4)$$

F is the total count for all words on a particular feature. If, for convenience, we think of vertical tags as *words* and horizontal tags as *features*, then F arises from summing vertically all values of a given (column) vector.

$$W = \sum_{j=1}^m c_{wj} \quad (4.5)$$

W is the total count for all features on a particular word. In this case, it arises from summing horizontally all values for a tag (row) vector.

$$P(w, f) = \frac{c_{wf}}{N} \quad (4.6)$$

$P(w, f)$ is the joint probability of a word-feature pair occurring together in the corpus (single cell in the matrix).

$$P(w) = \frac{W}{N} \quad (4.7)$$

$P(w)$ is the marginal probability of a word occurring in the corpus.

$$P(f) = \frac{F}{N} \quad (4.8)$$

$P(f)$ is the marginal probability of a feature occurring in the corpus.

Therefore, PMI is equal to:

$$PMI_{wf} = \frac{P(w, f)}{P(w) \times P(f)} \quad (4.9)$$

However, since PMI tends to overestimate values of less frequent observations, Pantel and Ravichandran (ibid.) suggest a discounting factor that will be multiplied with each PMI_{wf} :

$$\frac{c_{wf}}{c_{wf} + 1} \times \frac{\min(W, F)}{\min(W, F) + 1} \quad (4.10)$$

where c_{wf} is the number of times the word w appears with feature f . In the rest of this chapter, I use $PMI(w, f)$ (e.g. $PMI(\textit{art}, \textit{history})$), which is equal the natural logarithm (base e) of Pantel and Ravichandran's discounted version of pointwise mutual information:

$$\begin{aligned} PMI(w, f) &= \log \left(\frac{P(w, f)}{P(w) \times P(f)} \times \frac{c_{wf}}{c_{wf} + 1} \times \frac{\min(W, F)}{\min(W, F) + 1} \right) \\ &= \log(c_{wf}) - \log(W) - \log(F) + \log(N) + \\ &\quad \log(c_{wf}) - \log(c_{wf} + 1) + \\ &\quad \log(\min(W, F)) - \log(\min(W, F) + 1) \end{aligned} \quad (4.11)$$

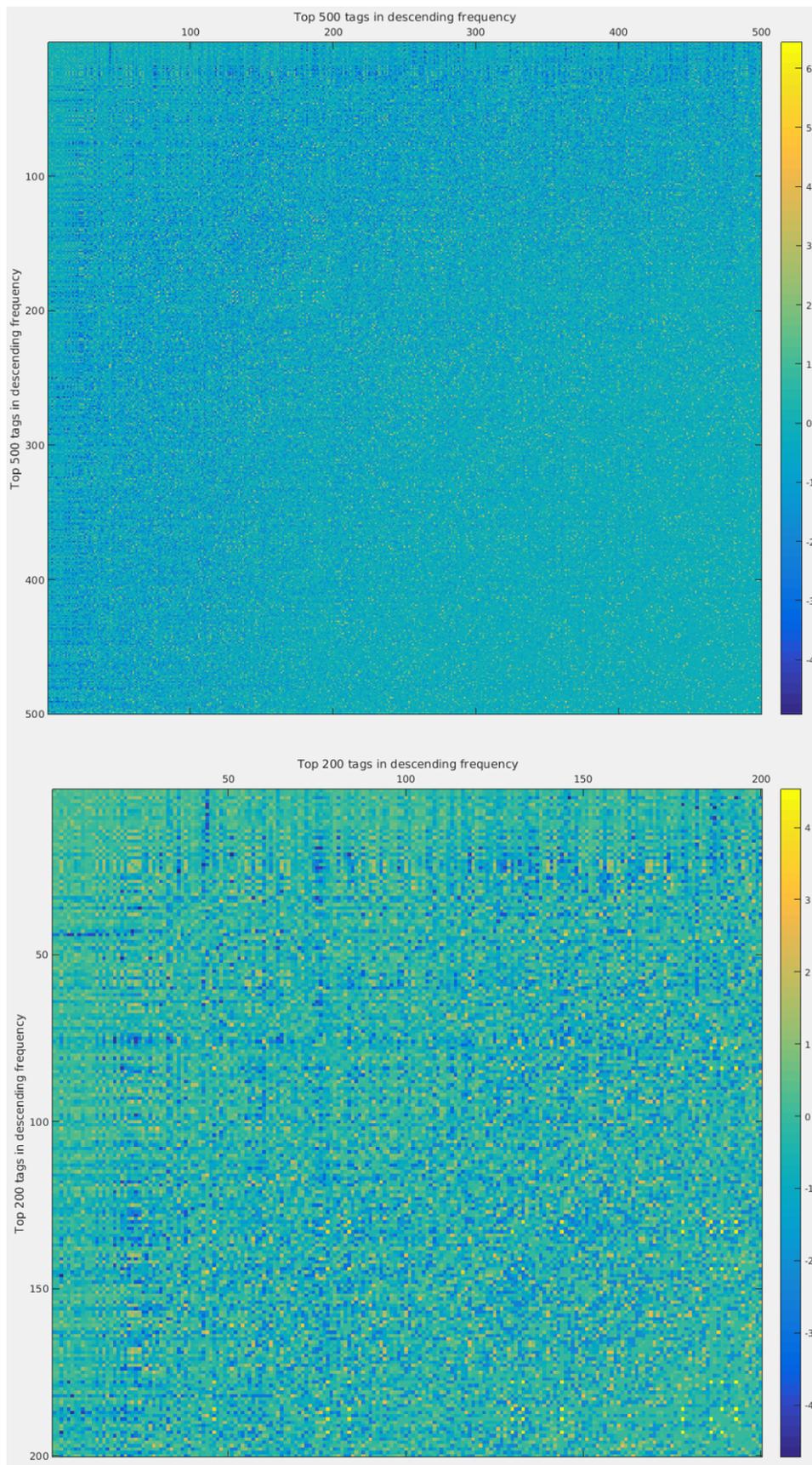


Figure 4.2: Tag PMI matrices

Figure 4.2 visualises the previous 2,382-by-2,382 matrix when using PMI instead of raw frequencies. In the figure, both x and y axis represent tags sorted from the most to the least frequent. The two matrices are portions of the graph illustrating the PMI values of the 500 and 200 most frequent tags respectively. We can see that PMI values of different magnitudes are spread evenly throughout the graph. For example, high values do not concentrate on pairs of frequent tags, as observed in the case of raw co-occurrence counts. A pair of tags that co-occur in the same tag cloud considerably more often than expected by chance (i.e. when $P(tag1, tag2)$ is much larger than $P(tag1) \times P(tag2)$) would have high PMI values. Such values, represented with yellow in the figure, can be seen even in the bottom right of each matrix, where the PMI values between lower rank tags are visualised. Likewise, low PMI values (represented by dark blue in the graph) can be seen between tags of various ranks. For example, the tag “paper” (rank 68) is associated with tags such as “author” (rank 1,681) or “folio” (rank 948) in tag clouds ($PMI(paper, author) = 3.2$, $PMI(paper, folio) = 3.07$) rather than tags such as “stone” (rank 45) or “ceramic” (rank 76) ($PMI(paper, stone) = -2.4$, $PMI(paper, ceramic) = -3.6$).⁷

A vector of PMI values (horizontal or vertical line in the graph) corresponding to each tag can provide a better representation of the tag’s syntagmatic relations with other tags than a vector of raw co-occurrence count does. Combinatorial restrictions that underlie how tags combine in tag clouds can be better represented by a measure that encodes some notion of inherent association between tags. In the following experiments, a syntagmatic vector is equivalent to a vector of PMI values.

4.2.2 Syntagmatic relations in folksonomy and natural language

The next step is to investigate whether the existence of idiosyncratic tag pairs provides evidence for lexical cohesion within tag clouds. To answer this question, I compared the PMI vectors of tags as they co-occur in tag clouds (“folksonomy-based syntagmatic vectors”; henceforth FSVs) with the PMI vectors of the same tags when they co-occur as words in general English text (“text-based syntagmatic vectors”; henceforth TSVs). High similarity between the FSVs of tags and the TSVs of the same tags would indicate that tag pairs that are associated in tag clouds are also associated in text. Since text is very likely to contain cohesive discourse, this would be evidence that such tag pairs help make tag clouds themselves cohesive entities. I, therefore, created a second matrix with PMI values for the same tags as the ones in Section 4.2.1, but this time the weights were learnt from Wikiwoods. The process of creating a PMI matrix from Wikiwoods was as follows:

1. **setting a threshold:** discarding all the tags that appear less than 20 times in Wikiwoods. After this step, word types in Wikiwoods were reduced from 5,051,015 to 412,410.
2. **reducing the vocabulary:** restricting the word types to those that appear as tags in Steve, since the context provided by the rest of the words was irrelevant for this task. After this step the vocabulary consisted of 2,059 types.
3. **creating a Wikiwoods matrix:** initialising a 2,059 by 2,059 empty matrix that would contain PMI values of word pairs that co-occur in sentences. Each cell $i - j$ (and its mirror cell $j - i$) was filled with the PMI value of i^{th} and the j^{th} word in

⁷Negative PMI values express dissociation, that is lower-than-chance co-occurrence.

the matrix, sorted by the count of their tag equivalents in Steve. For instance, as before, the cell 0–1 would hold the PMI of words “black” and “white” in Wikiwoods sentences, since these are the top two tags in Steve and they are also part of the Wikiwoods vocabulary after the reductions in the previous steps.

4. **shortening the Steve matrix:** removing vectors from the 2,307 Steve PMI matrix that represent words not found in the reduced Wikiwoods vocabulary. For each word not found in the 2,059-row square Wikipedia matrix, the relevant row (and the identical column) vector were removed from the Steve matrix. After this stage, the Steve matrix had dimensions 2,059 by 2,059, mirroring the structure of the Wikiwoods matrix.

Once the two equally-sized matrices (‘Steve’ and ‘Wikiwoods’) had been constructed, I performed a pairwise comparison of their vectors using cosine similarity. Since the indices of the two matrices represented the very same terms (e.g. cell 0 – 2 contained a PMI value for the word pair “black”-“tree” in either matrix), each row vector i in Steve was compared with each row vector i in Wikiwoods,⁸ resulting in a cosine similarity value within the range $[0, 1]$. Ultimately, a column vector of similarity scores was obtained. Each one of the values in the final vector shows how a given tag’s syntagmatic relations in Steve compare to the syntagmatic relations of the same term within Wikiwoods. The values of the vector are normally distributed and have a mean of 0.27. But what does this tell us? Is an average similarity of 0.27 between the syntagmatic vectors of Steve (S) and Wikiwoods (W) enough to support the claim that tags in tag clouds combine like words do in coherent text? To answer this question, we need to measure whether the syntagmatic vectors of a baseline folksonomy (B), say, one with random allocation of tags in tag clouds, have significantly lower average similarity with the vectors of Wikiwoods. In other words, if we measure a higher means of S - W vector similarities compared to the means of B - W similarities, then we can conclude that S is closer to W than B is. In turn, such a result would be an indication that the way tags are allocated in folksonomy tag clouds is closer to the way words are arranged in coherent text than randomly allocated tags are.

The baseline folksonomy which I created to make the above similarity scores more interpretable, is called ‘Semi-Random Steve’. This folksonomy contains **i)** the same number of resources as Steve, **ii)** tag clouds of the same capacity (i.e. number of tag tokens in a tag cloud) and **iii)** an identical frequency distribution of tags (as seen in Figure 3.2, p. 33), random allocation of tags in tag clouds. The process of creating Semi-Random Steve was as follows:

1. **emptying original tag clouds:** Tags from the 33,948 tag clouds of the Steve folksonomy were removed and tag cloud capacity was recorded. For instance, a tag cloud like {“art”: 1, “cubism”: 3, “woman”: 8, “mirror”: 5}, where “art” occurs once, “cubism” occurs three times and so on, has capacity of $1 + 3 + 8 + 5 = 17$ when empty.
2. **drawing tag tokens from the Steve distribution:** Tag *tokens* were ‘popped’ one-by-one from the head of the original (zipfian) tag distribution. The first 9,742 tokens popped were of the tag type “black”, the next 7,289 tokens were of the tag type “white” and so on.

⁸The same could be done with column vectors, since the matrix is symmetrical.

3. **randomly allocating popped tag token to a tag cloud:** For each popped tag token, a tag cloud was randomly chosen as a host. The token was stored in the tag cloud and the tag cloud’s capacity was reduced by one. If the tag cloud became saturated, then it was left aside and was no longer available to host new tag tokens. A saturated tag cloud of capacity 12 could look like this: “grass”, “grass”, “armchair”, “sculpture”, “black”, “black”, “white”, “candle”, “tree”, “figure”, “telescope”, “romantic”, or, alternatively this: {“grass”: 2, “armchair”: 1, “sculpture”: 1, “black”: 2, “white”: 1, “candle”: 1, “tree”: 1, “figure”: 1, “telescope”: 1, “romantic”: 1}
4. **creating semi-random folksonomy:** After all tag tokens had been allocated to tag clouds, the result was a baseline folksonomy with *random* combinatorial (co-occurrence) relations between tags in a tag cloud but *preserved* number of resources (and, by extension, number of tag clouds), capacity of tag clouds and distribution of tag tokens in the entire folksonomy. This allowed for the creation of a folksonomy that was as similar to Steve as possible, with only combinatorial restrictions between tags being randomised.

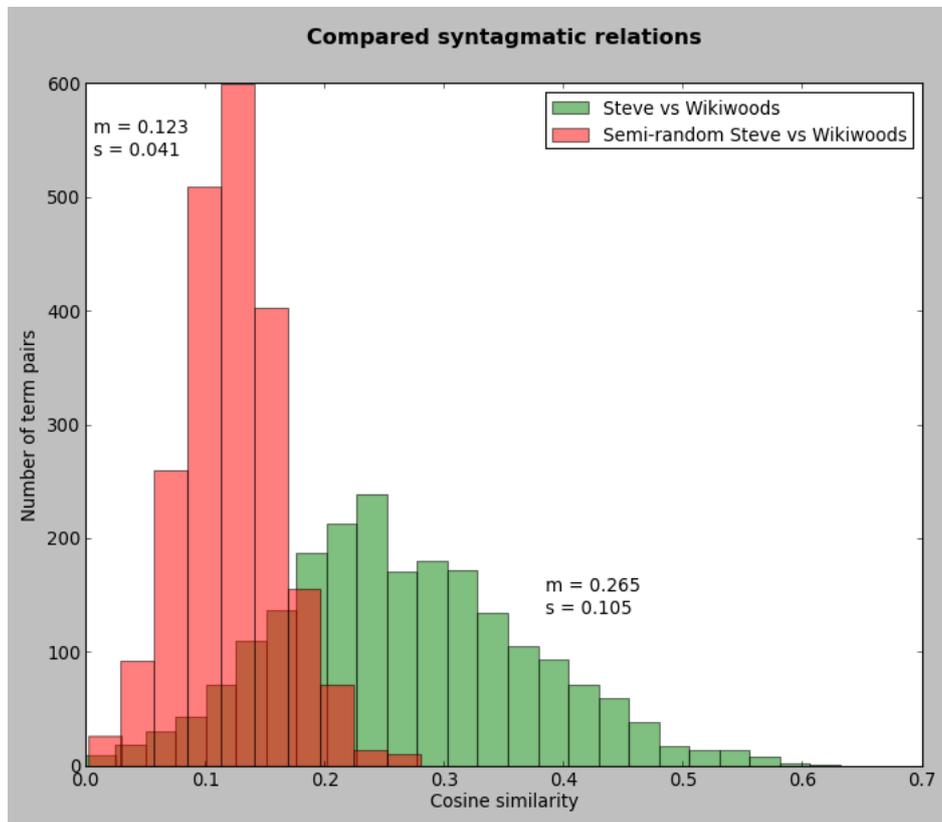


Figure 4.3: Compared syntagmatic relations between Wikiwoods and two tagsets

After creating the baseline, I created a PMI matrix from the Semi-Random Steve folksonomy, using the same tags that I had used to populate the reduced Steve and the Wikiwoods matrix. Thus, the Semi-Random Steve matrix had dimensions 2,059 by 2,059, mirroring the structure of the other two matrices. Semi-Random Steve was compared to Wikiwoods in the same way that Steve was, resulting in a 2,059-length vector of similarities. The values in the vector followed a normal distribution with mean similarity

0.12. The ‘Semi-Random Steve vs. Wikiwoods’ and the ‘Steve vs. Wikiwoods’ similarity distributions were plotted against each other (Figure 4.3).

Significance testing

To assess whether Steve tags combine like natural language words better than, worse than or as well as randomly assembled tags do, it is necessary to measure the probability that the two corresponding distributions of similarities (‘Steve vs. Wikiwoods’ and ‘Semi-Random Steve vs. Wikiwoods’) could be samples of the same underlying population. To obtain this probability, one can perform significance testing on the means of the two samples: assuming that the samples are indeed drawn from the same population (null hypothesis), how likely is it that the difference (Δ) between the means (m) of two randomly drawn samples is at least as extreme (two-tailed test) or as large/small (one-tailed test) as the one in the two samples of interest (in this case $\Delta = 0.27 - 0.12 = 0.15$). If the recorded difference (here 0.15) is highly unlikely to have occurred by random sampling from a common population, then one can conclude that the two samples are significantly different, that is they are drawn from two different underlying populations.

In order to avoid making normality assumptions for any of the distributions in question, I performed non-parametric significance testing based on Monte Carlo simulations, which have been used in NLP by Berg-Kirkpatrick et al. (2012), Yeh (2000) and Riezler and Maxwell (2005). In particular, I used *approximate randomisation* (Noreen, 1989), which simulates the underlying population of the null hypothesis by aggregating the individual values of both samples and shuffling them. Two samples of the desired size can be randomly drawn from this population and their means can be measured. The process of creating a population and drawing samples from it is repeated a number of times and the difference in the means is measured each time. As an example, if the process runs 1,000 times and the difference in the means of the randomly drawn samples is greater than the one in the samples of interest in only 5 of the runs, then one can conclude that the two samples have 0.05% probability of occurring by chance. This would mean that the difference between the two means is significant at a 99.5% confidence level, typically enough to reject the null hypothesis.⁹

Using approximate randomisation, I measured the difference between the means of the two distributions of similarities (‘Steve vs. Wikiwoods’ and ‘Semi-Random Steve vs. Wikiwoods’) and found it to be significant at a higher than 99.99% confidence level.

This result shows that tags assigned to folksonomy images obey similar combinatorial rules to those of their word counterparts in text, significantly more than randomly assigned tags do. If the way tags cluster together to form a tag cloud were random, their syntagmatic relations would have little similarity with those of text. However, the above experiments indicated that tags might combine in a tag cloud similarly to how they would combine in text. This implies that, at least with respect to word combinations, tag clouds are cohesive entities, like texts typically are. Cohesion itself is good predictor of coherence, which implies that tags in tag clouds can tell a story together and are not simply a list of unrelated keywords.

⁹Another popular simulation-based statistical test is called ‘bootstrap’ (Efron and Tibshirani, 1993). The difference lies in the fact that bootstrap re-samples with replacement based on the idea that all sample draws should be independent.

4.3 Tag similarity

As mentioned earlier in this chapter, co-occurrence patterns represent syntagmatic relations while similarity patterns, that arise by comparing co-occurrence vectors, represent paradigmatic relations. Having shown that syntagmatic relations in folksonomy resemble those in text, the natural next step is to construct paradigmatic (similarity) vectors and make further observations.

4.3.1 Similarity vectors

To construct a similarity matrix for Steve tags, where cell i, j indicates the similarity of tag i with tag j , I used the structure of the 2,059 by 2,059 PMI matrix (either from Wikiwoods or from reduced Steve; see §4.2.1) as a basis. Then I followed the steps below:

1. **creating empty matrix:** initialising an empty matrix that mirrors the structure of the Steve PMI matrix (i.e. has the same dimensions and each vector represents the same tags as in the PMI matrix).
2. **comparing vectors:** acquiring a cosine similarity score for each pair of words by comparing their vectors in the PMI matrix
3. **populating matrix:** inserting similarity scores to the right cells; e.g. the cosine similarity obtained by comparing vectors i and j of the PMI matrix is inserted in position i, j of the similarity matrix

Each vector in the similarity matrix contains paradigmatic relations for a given tag, which are estimates of the extent to which other tags can replace the tag of interest in a tag cloud.

4.3.2 Paradigmatic relations in folksonomy and natural language

While syntagmatic relations encode *constraints* on word (or tag) combination, paradigmatic relations encode *options*, or alternatives, for fulfilling a particular role (function) in syntagms. As already explained, the syntagmatic (PMI) vectors created for the Steve corpus can provide an explanation why a tag cloud such as {"salad": 6, "green": 5, "eating": 5, "food": 4, "people": 4, "healthy": 4, "yummy": 3, "cutlery": 3, "table": 1} seems more cohesive than a tag cloud like {"happiness": 8, "gloomy": 7, "driving": 5, "cake": 5, "pain": 4, "paper": 3, "polyurethane": 3, "grabbing": 3, "air": 2}. In other words, while PMI vectors tell us which tags are and which tags are not good *company* for a tag of interest (although the values are not binary), similarity vectors tell us which tags are and which tags are not good *alternatives* for a particular function. For instance, in the first tag cloud above, if we replace "cutlery" with "tablecloth", the overall cohesion of the tag cloud will not change noticeably, but if we replace "cutlery" with "keyboard", the tag cloud might become less cohesive. As mentioned earlier (§4.1.1), paradigmatic relations are often examined in the context of paradigmatic sets, that is sets of options which *are* inter-substitutable in a particular role.

To create paradigmatic sets for a particular tag t , we can take the vector which contains the tag's similarities with other tags, sort the the tags in the vector from the most to the least similar to t and select the top N tags that can replace t in a tag cloud without

violating combinatorial constraints (i.e. without affecting cohesion). Paradigmatic sets were constructed for Steve. Two examples (paradigms for tags “farmer” and “guitar” with $N = 20$) can be seen below (brackets contain similarity scores):

farmer: farm (0.53), agriculture (0.50), barn (0.47), field (0.47), harvest (0.47), rural (0.42), pastoral (0.41), crop (0.39), cows (0.38), pasture (0.37), wheat (0.37), peasant (0.36), countryside (0.36), sheep (0.36), farmland (0.34), autumn (0.33), labor (0.33), livestock (0.31), cart (0.30), valley (0.29)

guitar: instrument (0.44), musician (0.38), singing (0.37), strings (0.36), flute (0.34), playing (0.32), piano (0.32), instruments (0.26), drum (0.24), play (0.24), party (0.23), dancing (0.22), longneck (0.22), caricature (0.19), inlay (0.20), boys (0.19), inlaid (0.18), clock (0.18), mustache (0.17), laughing (0.17)

These paradigmatic sets contain tags that are intuitively similar to each other, hence inter-substitutable in a tag cloud. The above sets are equivalent to what Lin (1998) calls ‘thesaurus entries’, which were shown to resemble those in real thesauri (e.g. Roget’s Thesaurus). Paradigmatic sets were created not only from the Steve PMI matrix but also from the Wikiwoods PMI matrix (both of size 2,059; see above). An example (i.e. paradigm for teapot in Steve and in Wikiwoods) can be seen below. Italics represent terms that occur in both paradigmatic sets. The two sets are intuitively similar to each other.

teapot [STEVE]: *tea* (0.51), *saucer* (0.49), *teacup* (0.48), *kettle* (0.46), *pitcher* (0.46), *spoon* (0.41), *cup* (0.40), *jug* (0.40), *tray* (0.38), *lid* (0.35), *handle* (0.334), *urn* (0.29), *polished* (0.29), *useful* (0.29), *breakable* (0.28), *porcelain* (0.28), *container* (0.27), *pot* (0.26), *base* (0.25), *jar* (0.24)

teapot [WIKIWOODS]: *tea* (0.39), *pot* (0.38), *teacup* (0.37), *saucer* (0.34), *earthenware* (0.33), *lid* (0.33), *porcelain* (0.31), *container* (0.31), *spoon* (0.30), *jug* (0.30), *tray* (0.30), *kettle* (0.30), *bottle* (0.28), *jar* (0.28), *drink* (0.28), *clay* (0.27), *dome* (0.26), *glaze* (0.26), *handles* (0.25), *handle* (0.25)

The next step is to measure the extent to which paradigmatic sets of tags in Steve resemble those of their word counterparts in Wikiwoods. To achieve this, I compared all pairs of paradigmatic sets, such as those for “teapot” above, using Lin’s (1998) equation. The two sets for a particular term can be represented as:

$$\begin{aligned} w &: w_1 (s_1), w_2 (s_2), \dots, w_N (s_N) \\ w' &: w'_1 (s'_1), w'_2 (s'_2), \dots, w'_N (s'_N) \end{aligned} \tag{4.12}$$

where w is a tag from Steve and w' is an identical word from Wikiwoods; w_1 is the tag most similar to w with similarity s_1 , w_2 is the second most similar tag and so on. N stands for the size of thesaurus entries, which I had set to 20. Similarity between terms w and w' is defined as:

$$\frac{\sum_{w_i=w'_j} s_i s'_j}{\sqrt{(\sum_{i=1}^N s_i^2)(\sum_{j=1}^N s'_j{}^2)}} \quad (4.13)$$

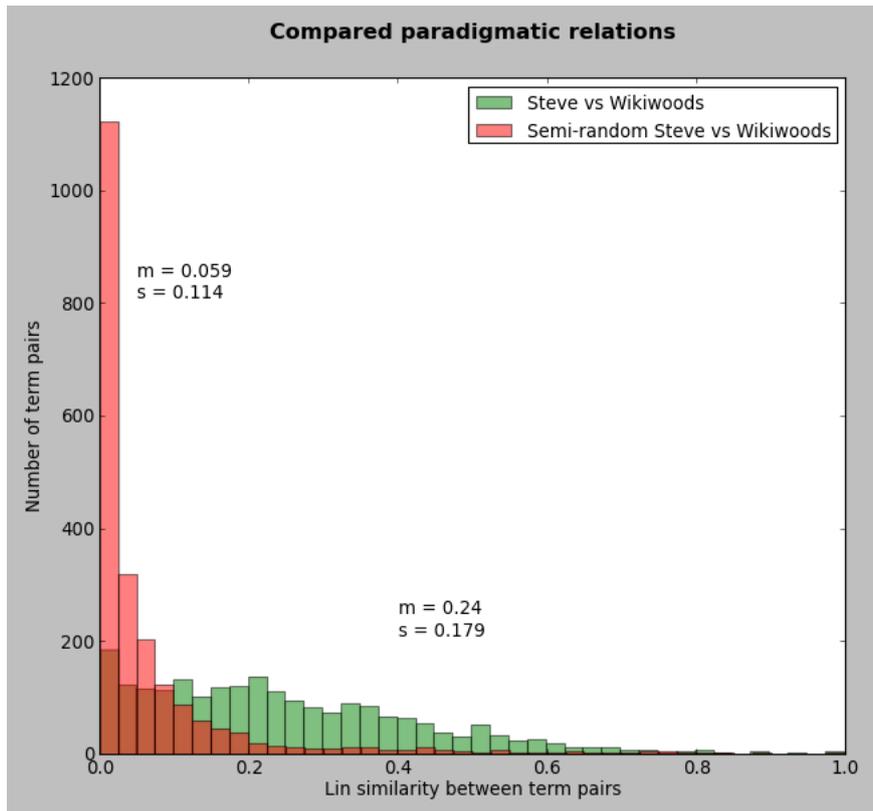


Figure 4.4: Compared paradigmatic relations between Wikiwoods and two tagsets (Steve and Semi-random Steve)

Using the above similarity measure, I constructed a vector of length 2,059, that contains similarities between the paradigmatic sets of Steve and Wikiwoods for each one of the 2,059 tags of interest. The mean similarity in this vector was 0.24. To make this number easier to interpret, I also constructed a vector of the same length, that contained similarities between the paradigmatic sets of Semi-Random Steve and Wikiwoods. This baseline similarity vector had mean 0.059 and a distribution that was highly skewed to the left because of the large number of zero similarities. The probability distribution of both vectors is visualised in Figure 4.4. Using the same significance test as before (§4.2.2), I found that the difference between the means is statistically significant at a higher than 99.99% confidence level.

What this result shows is that paradigms in folksonomy are similar to those in natural language, which is an indication that tag clouds and text have slots for similar ‘roles’, or units of meaning, during the construction of particular syntagms.

4.4 Summary

In this chapter, I compared the co-occurrence patterns of tags in tag clouds with those of words in sentences. I claimed that comparing such patterns is equivalent to comparing the syntagmatic and paradigmatic relations that hold between tags with those holding between words. I showed that such relations manifest between tag pairs similarly to how they manifest between words in coherent sentences. This provides an indication that tag clouds themselves are coherent entities.

Chapter 5

A parallel corpus of tags and text

In the previous chapter, I showed that the way tags combine to annotate an image resembles the way the same tags would combine as words in coherent text. The next step is to determine the extent to which related tag pairs are governed by underlying semantic relations. For instance, in natural language text, “food” and “plate” often appear in the same sentence because they belong to the same domain. But in a subset of those sentences, they may be connected with relationships such as “there is some food on some plate”, “this place is for food” and so on. Could it be that particular tags co-occurring in a tag cloud are traces of similar semantic associations made by the users? To provide evidence that particular tags used for a given image are connected with implicit semantic relations, it is interesting to examine what users *would have said* if they were to describe an image with fully fledged text. Assuming that a user has the same understanding of the image while tagging it and while describing it with text, then textual descriptions of a resource can provide a ‘window’ into the user’s thoughts. To test the extent to which such relations exist, as well as the nature of those that do exist, I compiled a corpus of 1,090 parallel tags-text annotations submitted by 218 participants, each one of whom both tagged and described the same set of images. This is, to my knowledge, the first attempt to construct a parallel corpus of tags and textual descriptions with respect to particular resources.

The purpose of this chapter is to describe the construction of the parallel corpus. The first Section outlines the specific objectives of the corpus compilation experiment. Section 5.2 describes the experimental design that formed the basis of the experiment, the details of which are given in Sections 5.3 (pilot experiments) and 5.4 (final experiment). Section 5.5 describes the completed corpus and discusses the quality of the collected data. Finally, Section 5.6 provides an overview of existing tagging corpora, explaining why they were inadequate for this task.

5.1 Objectives

The objectives of the corpus compilation experiment can be summarised as follows:

1. to provide experimental *confirmation* for the hypothesis that implicit relations can hold between tag pairs annotating a particular resource. More specifically, the goal was to measure the *extent* to which such relations may hold (e.g. How many of the possible tag pairs in a tag cloud are semantically linked? Put differently, what

proportion of possible pairs of tags can we plausibly extract relations for?), as well as the *nature* of relations that were found (e.g. How are relations lexicalised in the text? Do tags follow the same order in an individual user’s tag annotation and her/his textual description of an image?).

2. to build a folksonomy enriched with natural language descriptions, which are provided by each user for each picture along with the respective tags. In other words, the aim is to produce a more informative folksonomy, consisting of four sets (Tags, Users, Resources and Descriptions) and a quaternary relation that holds between the sets (see §1.2 for a formal definition of folksonomy). This enriched folksonomy can act as a parallel corpus of annotations (tags + textual description) for each image by each user and can provide valuable information as to what the user was intending to ‘say’ when labelling an image with particular keywords. These tagging data and natural language descriptions are acquired under a controlled environment, where the tagging purpose, the description purpose as well as the interface are the same for all users. With such a folksonomy, one can also test how the tag-description interaction differs if tags and descriptions are examined *collectively* (i.e. from all users) or within *individual* user annotations. Finally, this corpus can constitute a useful language resource for further theoretical research (e.g. on tagging behaviour or visual description and how they related to demographic characteristics of the participants, some of which were collected).
3. to supply the tools and methodology necessary for future construction of larger-scale parallel corpora that can aid resource-specific relation extraction.

5.2 Experimental design

In order to construct the parallel corpus, I designed an experiment in which participants perform **i)** a *tagging task*, under conditions similar to the ones in collaborative tagging systems and **ii)** a *description task*, whose instructions outline the purpose of textual description but do not predispose users towards a particular writing style. Below is a discussion of the relevant variables and conditions, participant groups and visual stimuli.

5.2.1 Variables and conditions

The *independent* variable of the experiment (i.e. the one that the researcher can manipulate) is the way an image is annotated; a categorical variable which can take two values: tagging and describing. A number of *dependent* variables can be tested (e.g. Does the probability distribution of tokens differ between tag annotations and textual descriptions?), although such questions were not specified in advance. The reason was that the primary purpose of the experiment is to collect parallel data in order to examine the existence of underlying inter-tag relations and not to investigate cause and effect relationships. Two *extraneous* variables (i.e. ones that might affect the result but cannot be changed by the experimenter) were controlled for. These were : **i)** whether or not a participant is a native speaker of English (since both tasks had to be performed in English), **ii)** whether or not a participant has had previous experience with online tagging.

5.2.2 Experimental and control group

For this experiment, it was important that each subject should participate in both conditions of the experiment, that is tagging and describing, so that one’s tags could be compared with their own textual descriptions. Hence, I adopted a *repeated measures design*, in which all participants are used to test both conditions of the independent variable. The two tasks, tagging and textual description, were performed by the same users in two separate phases (Phase One and Phase Two). Between the two phases there was a two-week gap, which was thought to be long enough to minimize repetition bias and short enough to ensure that participants had the same perspective on a given image as in the previous phase.

To eliminate order effects, I used *counterbalancing*, whereby half of the subjects perform the tasks in one order (e.g. tagging in Phase One and describing in Phase Two) while the other half perform the tasks in the reverse order (i.e. first describing and then tagging). The idea behind counterbalancing is that any error arising from order effects will be spread equally across the participants, so it will not affect subsequent measurements.

5.2.3 Stimuli

Five images were chosen from the Steve dataset to be presented to participants (see Figure 5.1). Highly tagged images were preferred because this allows the original tags to be later compared to the tags acquired from this experiment. Among the top 70 most tagged images (i.e. the ones with the most tag tokens), five were hand-picked according to two criteria: **i**) complexity of image (simple images such a single sculpture against a white background were dis-preferred because of the limited possibilities they offer in term of possible tags and underlying relations) and **ii**) diversity (the five images chosen were reasonably different to each other in terms of themes, style and clarity of the messages conveyed).

During the experiment, the original tags, official metadata and previous participants’ tags were not available to the subjects, in an attempt to elicit tags and descriptions that are as unbiased as possible. Another reason was to explore if the tags provided by all users follow a power-law distribution even without users imitating each other’s tagging behaviour (see discussion in Section 3.1).



(a) “House in Provence” (1885) by Paul Cézanne



(b) “The Two Sisters” (1894) by Georges Lemmen



(c) “Angel of Resurrection” (1904) by Louis C. Tiffany



(d) “Torso” (1959) by David Park



(e) “Moulin Rouge: La Goulue” (1891) by Henri de Toulouse-Lautrec

Figure 5.1: Visual Stimuli

5.3 Piloting

In order to assess the feasibility and usefulness of the experiment, identify possible problem areas and decide on the appropriate terminology for instructions, three small-scale pilot experiments were conducted. At the end of each pilot study, participants were asked to discuss their overall experience with the experiment and provide feedback on the clarity of the instructions. Each pilot study was undertaken by a different group of participants, all of whom were University of Cambridge postgraduate students unfamiliar with the objectives of the experiment.

5.3.1 Pilot study 1 (October 2012)

The first pilot experiment was conducted in an exploratory fashion, with no fixed instructions, allowing participants to ask for clarification but at the same time avoiding examples or anything that could bias their responses.

The five images in Figure 5.1 were printed in five separate sheets and given to each participant as a shuffled stack. On the first phase, participants were asked to write down on a separate sheet some “tags” or “keywords” that they would use to label the image in order to find it easily in the future. During this phase, participants were provided with alternative instructions if the task was still unclear to them (for instance, “descriptive keywords”, “label the image”). After the end of the tagging phase, responses were collected and hidden from the users, image prints were shuffled and given back to the participants. On the second phase, participants were asked to provide some “text” or “proper description” of the images. They were encouraged to ask for alternative instructions but overly

specific instructions were avoided. This pilot was taken by three University of Cambridge students. In particular:

- *Participant 1*: native English speaker, no tagging experience (full responses on table A.1, page 163)
- *Participant 2*: non-native English speaker, no tagging experience (full responses on table A.2, page 164)
- *Participant 3*: non-native English speaker, no tagging experience (full responses on table A.3, page 164)

Feedback The users' comments focused on the terminology used for instructions. Below are some of the terms used in the instructions and how they were received by the participants.

TAGGING

- “tagging”: not clearly defined
- “search terms”, “descriptive terms”, “notes”: easy to understand
- “terms you would use if you were to search for it on the internet”: easy to understand
- “terms you would use if you were to search for it on the the computer”: confusing (Participant 2 interpreted this as ‘file directories’; see this person’s responses on table A.2, page 164)
- “write whatever you think is relevant”: too vague

DESCRIPTIONS

- “descriptions”, “proper descriptions”, “text” and “notes”: too vague
- “some text that describes the picture”: unclear
- “describe and explain the picture”: unclear

Action Taken The above feedback was taken into account for the development of the web interface that was used in the second pilot experiment. Vague terms were eliminated from the instructions and both tagging and descriptions were designed to proceed with a real-life scenario in mind.

5.3.2 Pilot study 2 (November 2012)

This pilot experiment was completed by three participants. The first two were asked to provide tags for the five images and the third one was asked to provide textual descriptions. The participants were presented with a web interface (purpose-built script running on a laptop's localhost) which welcomed them, asked them to provide their name and email address and described a scenario. Users were asked to imagine that they are saving the images-to-follow on a personal online image gallery and they had to label them with tags that would help them retrieve the pictures later. On the next pages, the five images (Figure 5.1) were presented one after the other in a fixed order for all subjects. Each image was presented in its own page, along with a text field where participants could type in at most 30 tags. Participants were asked to write each tag on a separate line and then click 'Next' to proceed to the next picture. The third participant was asked to provide a description for each one of the above images (in the above order). The descriptions had to be at most 500 characters long. As a usage scenario, the participant was asked to imagine that they were describing the picture to a person with impaired vision. In this pilot, participants were not allowed to ask for clarification but were encouraged to comment on their problems at the end of the experiment. All three participants were University of Cambridge students:

- *Participant 1*: native English speaker, no tagging experience (full responses on table A.4, page 166)
- *Participant 2*: native English speaker, no tagging experience (full responses on table A.5, page 166)
- *Participant 3*: non-native English speaker, some tagging experience (full responses on table A.6, page 167)

Feedback Below is a summary of the comments made by each participant:

Participants 1 and 2 (tagging)

- Separate fields for tags would be more helpful than the supplied textbox.
- It was not clear if tags could be more than one word long.¹
- Giving a reason for tagging (i.e. the scenario) was very useful.

Participant 3 (description)

- The interface was easy to use, clear and minimal (with no distractions).
- It would be more useful to ask for minimum (rather than just maximum) number of characters or number of lines.

¹This was intentionally left unclear.

Action Taken After this pilot experiment a few formatting changes were made to the web interface, fields were provided for tag entry, minimum required text was specified and usage scenarios were presented.

5.3.3 Pilot study 3 (January 2013)

This pilot was conducted in a way that would simulate the final experiment. The interface ran from a web server, so it was accessible from any computer over http, while pictures were shuffled so the order was unpredictable. A link was emailed to each one of the two participants. The first participant was asked to provide textual descriptions on the five pictures (Figure 5.1) and the second participant was asked to provide tags. The order of the images was randomised. No further instructions were given apart from the ones already in the interface and there was no contact between the participant and myself during the experiment. At the beginning of the experiment, participants were asked if they are native speakers of English and if they have had previous tagging experience. Both participants were University of Cambridge students:

- *Participant 1*: non-native, some tagging experience (full responses on table A.7, page 167)
- *Participant 2*: native English speaker, no tagging experience (full responses on table A.8, page 168)

Feedback The comments received by each participant are summarised below:

Participant 1 (description)

- When users are asked for textual descriptions, the question of whether they are familiar with tagging is irrelevant and confusing.

Participant 2 (tagging)

- Having a scenario for tagging was very helpful.
- The fact that the pictures are so different from one another is also helpful because it keeps the participant interested.
- Some parts of the instructions are not idiomatic English.
- It would be interesting to ask what the participant's mother tongue is and not just whether they are native English speakers or not.

Action Taken After this pilot experiment, the web interface was finalised. The main changes were:

- Participants were asked if they are familiar with tagging only before tagging pictures and not before providing textual descriptions, to avoid confusion. Originally, this question was asked on the first phase of the experiment, whether it was a tagging task or a description task. After this pilot, the question was asked before a tagging task, whether it was first phase or second phase.
- The language of the instructions was improved.

5.4 Final Experiment

5.4.1 Recruitment and data collection

The final experiment was completed by members of the University of Cambridge. The reason for limiting the sample to a university-internal audience was to avoid introducing too many demographic parameters that might inhibit the experimenter’s ability to make generalisations about the data collected. Methods like crowd-sourcing (e.g. with Amazon’s Mechanical Turk²) were avoided because they could introduce unnecessary noise, for instance, dishonest responses that are hard to filter out (Ipeirotis et al., 2010). Each participant who completed both phases of the experiment could participate in a draw for a £100 voucher. The maximum amount of time that completion of each task was expected to take was clearly stated. The experiment was advertised to the following groups of people based at the University of Cambridge: Natural Language and Information Processing group (Computer Laboratory), friends, Trinity Hall graduate community, Language Sciences Initiative, Graduate Union, newsletters for graduate and undergraduate communities in colleges (through secretaries or communications officers), Faculty of Modern & Medieval Languages and Faculty of English.

The full call-for-participants email can be seen in Appendix A.4.1, page 169. Participants could follow a web link included in the recruitment email. Those who completed Phase One were sent a personalised web link by email asking them to complete Phase Two. Upon completion of Phase Two, participants were entered for the prize draw.

During the experiment, care was taken to avoid collecting personal information that was not essential for the experiment (e.g. gender and age), and to clarify that participants have the right to abandon the experiment at any time if they wish so. It was also made clear that participation in the experiment is voluntary and not tied to academic obligations. This policy was approved by the Ethics Committee of the Computer Laboratory, University of Cambridge.

The data collection process took place from 8th February to 8th March 2013. Phase One was completed in the first seven days (8th-15th February 2013) and Phase Two was conducted after a two-week gap (1st-8th March 2013).

5.4.2 Task presentation

Both the tagging task and the description task were performed in the context of real-life scenarios, which were informative enough to be clearly understood and open-ended enough to avoid prescribing a particular tagging or describing behaviour. Apart from the

²<https://www.mturk.com/mturk/welcome>

added motivation that a realistic scenario would provide to the user, it would also ensure that different participants are tagging or describing with the same purpose in mind.

Tagging task

Before starting the tagging task, subjects were given the following instructions:

Imagine that there is an art website which contains images of artworks; let's call it **www.my-personal-gallery.com**. This website allows you to register, choose your favourite art images and create a personal collection. In order to organise your images and be able to find them in the future, the website allows you to label them with keywords (**tags**). Now you will be shown **5 pictures**. Please provide tags for each one of them. You are free to type in anything as a tag as long as it helps you retrieve the picture from your collection later.

The above text instructs the participants to treat tagging as a tool for organising their personal data for the purpose of future retrieval, which is the main reason behind online tagging (see §1.1). Hence, this scenario is meant to elicit tags similar to those found in existing collaborative tagging systems.

During tagging, participants were presented with the five images (Figure 5.1), each on a separate webpage. Each image was accompanied by a list of text fields, which acted as slots for recording tags. Each tag had to be written on a separate field, which allowed for whitespaces to be used as word delimiters within multi-word tags. The first five fields (i.e. tags) were mandatory and another 15 fields were optional, offering slots for a maximum of 20 tags.

Description task

Before starting the description task, subjects were presented with the following scenario:

Imagine that you are in a bookshop holding a book in your hands. The book contains art images. Next to you there is a person with impaired vision and they are asking you to describe and explain to them what the pictures are about. Now you will be shown **5 pictures**. Please describe them to this person.

This scenario was meant to help users produce descriptions without forcing them to do too much guesswork. At the same time, it avoids imposing a particular format (e.g. “write a paragraph”) or a particular information content (e.g. “describe what you see”, “describe how the image makes you feel”).

During the description task each image was accompanied by a large text field, requiring “at least 3 lines”, which was a simplified way of asking participants for a minimum of

120 characters before they were allowed to proceed to the next page. The text field could accommodate a maximum of 500 characters.

5.4.3 Interface

Phase One

Participants could start Phase One of the experiment by following the URL address³ included in the recruitment email. This address redirected them to the task that they were supposed to perform first. The script logged details of each response and assigned a tagging or a description task to participants interchangeably by looking at the number of responses received on each task. A given participant was assigned to a task for which the fewest responses had been received (usually one response less). Once the task was chosen, the relevant page loaded and the experiment started. In Phase One, all participants were shown the following eight pages, one after the other:

- Welcome page
- Scenario page
- Image pages (×5)
- Thank You page

Welcome page Both versions of the welcome page (i.e. for those tagging and for those describing) started by thanking the participant and explaining that the experiment’s research objective is to study the “language people use to describe pictures”. After a brief and generic description of the task, the users were asked whether they are native speakers of English. They were also required to provide their email address, which was used as a contact for the second phase, and to acknowledge that they are committing to undertake Phase Two when requested. For those undertaking the *tagging* task, there is an additional question asking whether the participant has had previous experience with online tagging. After responding to all questions, participants could press a ‘Start’ button to proceed to the Scenario page. The Welcome page for each task can be seen in Figures A.1 (p. 172) and A.9 (p. 176).

Scenario page For the tagging task, this page displayed the image retrieval scenario described in §5.4.2 (p. 68). For the description task, the art bookshop scenario was displayed (also in §5.4.2). At the end of the instructions, there was a button titled ‘I’m ready to tag’ for participants tagging, or ‘I’m ready to describe’ for participants providing textual descriptions. Screenshots of the Scenario page can be found in Figure A.2 (p.172) for the tagging task and in Figure A.10 (p.176) for the description task.

At this stage, images were shuffled for each participant in a way that generates heterogeneous sequences, that is sequences with large enough edit distance from each other (see details in §A.4.2, p. 170). The rationale behind differentiating, as opposed to randomising, the order of images displayed was that heterogeneous sequences would be more representative of the complete set of permutations (i.e. $5! = 120$) in case the number of

³<http://www-dyn3.c1.cam.ac.uk/~tt309/experiment-lent-2013/code/phase1/index.cgi>

participants was low. Given the large number of participants that this experiment attracted, simply randomising the order of the images would have been adequate, although differentiating guaranteed that all permutations were used. After a sequence of images was selected, the participant was presented with five image pages, one after the other.

Image pages Each one of the image pages displayed an image on the left and a list of text fields (in the tagging task) or a text box (in the description task), as described in §5.4.2. After submitting their data, users had to press the ‘Next’ button to proceed to the next image. Screenshots for the tagging task can be seen in figures A.3, A.4, A.5, A.6 and A.7 (p.173-175). Likewise for the description task: figures A.11, A.12, A.13, A.14 and A.15 (p.177-179). When the last image was reached, the ‘Next’ button directed participants to the Thank You page.

Thank You page The final page of Phase One thanked the subjects for their time and promised that they will be contacted two weeks later for Phase Two. Screenshots of the Thank You page for the tagging task can be seen in Figure A.8 (p.175) and for the description task in Figure A.16 (p.179).

Phase Two

Subjects who had successfully completed Phase One were invited to participate in Phase Two through a personalised web link sent to their email address. The URL contained the participant’s email address as a parameter, so their responses could be matched with the ones they provided in Phase One. After responses had been matched, information about email addresses was destroyed and each participant was represented by a unique ID. In Phase Two, participants were asked to perform the task that they had not completed yet (i.e. describing for those that tagged in Phase One and vice versa). When clicking on the web link, participants were presented with the Welcome page of Phase Two.

Welcome page (Phase Two) In the Welcome page of Phase Two, participants were informed that they were going to see the same five pictures as in Phase One, but “possibly in a different order”. Those who had already completed tagging were now asked to provide textual descriptions. A screenshot of this page for the description task can be seen in Figure A.17 (p.180). Those who had already described the images with text were now asked to tag them. They are also asked if they have previous experience with tagging. The Welcome page for the tagging task of Phase Two can be seen in Figure A.19 (p.181).

Scenario page (Phase Two) The scenario page for each task was identical to the one presenting the same task in Phase One.

Image pages (Phase Two) The five image pages for each task had the same format as the ones presenting the same task in Phase One. The order of the images was very likely to be different from the one a participant had seen in Phase One.

Thank You page (Phase Two) The final page thanked the subjects for participating in both Phases and informed them that they would be entering the prize draw originally

advertised. This page had the same wording for both the tagging task (Figure A.20, p.181) and the description task (Figure A.18, p.180).

Table 5.1 shows in what ways the pages displayed during the experiment differed, either across tasks or across phases. As can be seen on the table, two equally-sized groups of participants (grey columns) performed the two tasks in a different order each. Scenario pages were the same for the tagging task (blue bullet) occurring in either phase; likewise for the description task (magenta bullet). The format of image pages was the same for the tagging task (blue star) regardless of phase and likewise for the description task (magenta star), although the images displayed were shuffled for every participant. Thank You pages were the same for Phase One (orange clubsuit) regardless of task and likewise for Phase Two (green clubsuit). Welcome pages were similar for Phase One regardless of task (and likewise for Phase Two).

	1/2 of participants	1/2 of participants
Phase One	Tagging task 1. Welcome page  2. Scenario page  3. Image pages (×5)  4. Thank You page 	Description task 1. Welcome page  2. Scenario page  3. Image pages (×5)  4. Thank You page 
	Description task 1. Welcome page  2. Scenario page  3. Image pages (×5)  4. Thank You page 	Tagging task 1. Welcome page  2. Scenario page  3. Image pages (×5)  4. Thank You page 
Phase Two		

Table 5.1: **Pages displayed during experiment**

Same colour-symbol combinations indicate (almost) identical page formats. Some pages differ across phases, while others differ across tasks.

5.5 Resulting corpus

Phase One of the experiment was completed by 267 people (134 providing tags and 133 providing descriptions), 219 of whom completed Phase Two. Data from the 48 participants who did not continue to the second phase was ignored during the construction of the parallel corpus. Phase One was structured in such a way that it could guarantee an (almost) equal number of participants assigned to each task (see §5.4.3); in this experiment, there was an odd number of participants, which explains why those who tagged

outnumbered those who described by one. However, Phase Two did not offer such guarantees because there was no way to predict which particular individuals would proceed to the next phase. It was expected that drop-out rate would be approximately equal for both categories of participants and, indeed, those who completed Phase Two having provided tags in Phase One (110 subjects) were almost as many as those who completed Phase Two having provided descriptions in Phase One (109 subjects). To keep the corpus balanced and eliminate potential order effects (see §5.2.2), a participant was randomly chosen among those who started the experiment with tagging, and her/his responses were discarded. This resulted in a corpus of 218 participants (109 tagging in the first phase and describing in the second phase, and 109 doing the reverse). Each one of these subjects had annotated five pictures with parallel tags-text data, thus, the final corpus contains 1,090 parallel annotations. Example data from the parallel corpus can be seen in Figure 5.2.

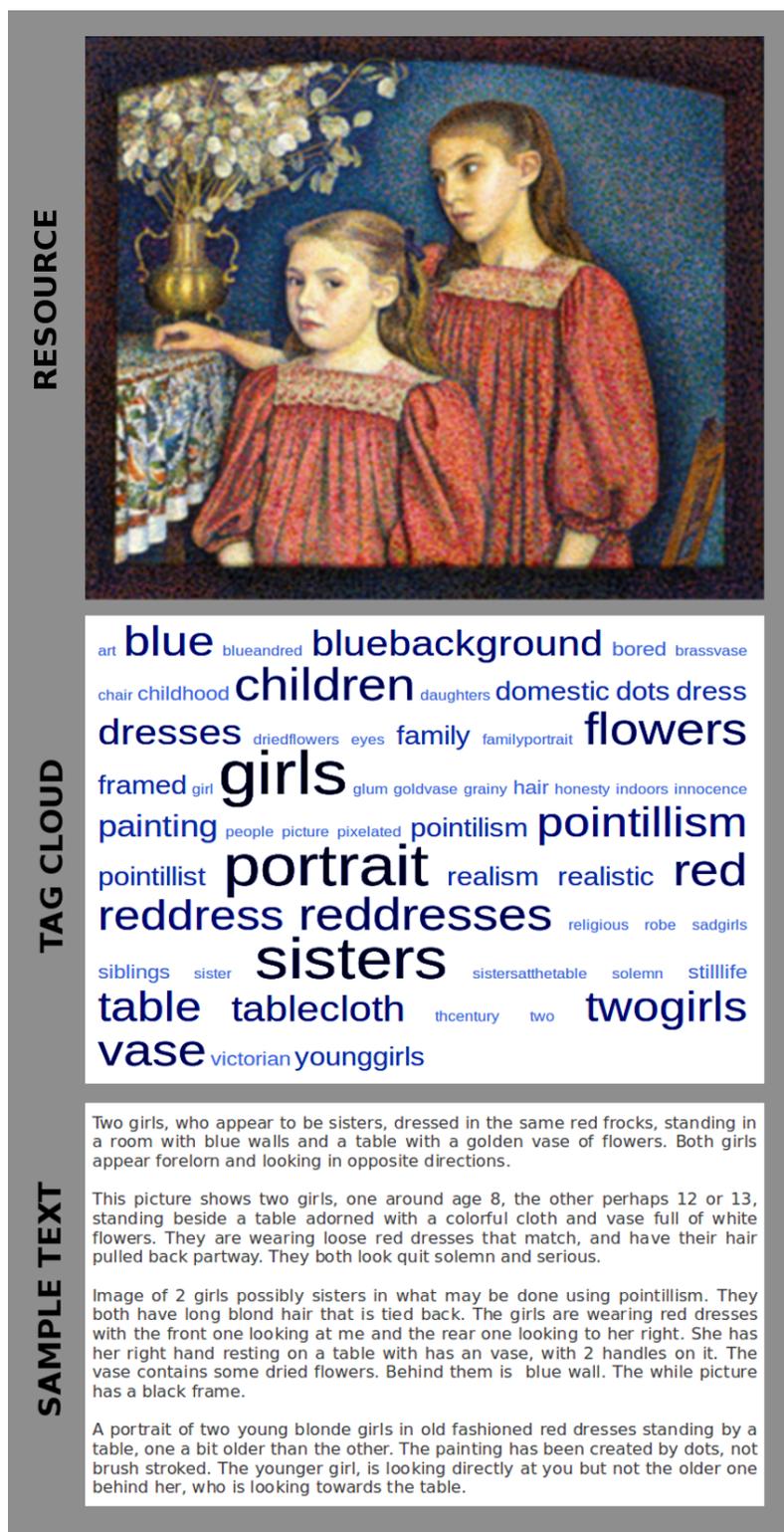


Figure 5.2: **Example parallel corpus data.** A tag cloud from the tags collected from all participants for image “Two Sisters” (with hapax legomena omitted because of space limitations) and a sample of descriptions (from four participants)

Format The data was saved in a *.csv* file, in which each line represents data from a different participant. Fields are separated by tab characters, text is delimited by * (asterisks) and tags are separated by | (vertical bars). The following information has been recorded for each participant, presented here with IDs for each field as they appear on the header line of the corpus:

<i>participant_id:</i>	e.g. "Participant_104"
<i>phase1:</i>	"TAGG" if the participant performed tagging in Phase One "DESC" otherwise
<i>native_speaker:</i>	"YES" if the participant is a native speaker of English "NO" otherwise
<i>familiar_with_tagging:</i>	"YES" if the participant has previous tagging experience "NO" otherwise
<i>house_D:</i>	e.g. "This is a house in the fields. The painting is pretty." as description of image in Figure 5.1a
<i>house_T:</i>	e.g. "house countryside painting by Cezanne pastoral" as four different tags for the image in Figure 5.1a
<i>moulin_D:</i>	textual description for the image in Figure 5.1e
<i>moulin_T:</i>	tags for the image in Figure 5.1e
<i>torso_D:</i>	textual description for the image in Figure 5.1d
<i>torso_T:</i>	tags for the image in Figure 5.1d
<i>angel_D:</i>	textual description for the image in Figure 5.1c
<i>angel_T:</i>	tags for the image in Figure 5.1c
<i>sisters_D:</i>	textual description for the image in Figure 5.1b
<i>sisters_T:</i>	tags for the image in Figure 5.1b

The full corpus is available on:

http://www.cl.cam.ac.uk/~tt309/parallel_corpus.csv.

Participant variables As seen in Section 5.2.1, there are two participant variables which could affect measurements: **i)** whether or not a subject is a native speaker of English (let us call this variable *Native*) and **ii)** whether or not a subject is familiar with tagging (let us call this *Experienced*). Both of these variables are binary, so they could be represented together in a 2-by-2 contingency table. Figure 5.3 presents the number

of people for which particular *Native* and *Experienced* values are true. For example, there are 40 participants who are not native speakers of English and, at the same time, are not familiar with tagging (i.e. *Native* = 0 and *Experienced* = 0). Among the 218 participants in the corpus, 162 (74.3%) are native speakers of English. With respect to tagging experience, 157 subjects (72% of 218) have never performed tagging prior to the experiment.

	<i>Native</i> = 0	<i>Native</i> = 1	Total <i>Native</i>
<i>Experienced</i> = 0	40	117	157
<i>Experienced</i> = 1	16	45	61
Total <i>Experienced</i>	56	162	

Table 5.3: **Contingency table for two participant variables.** Variable *Native* equals 1 when someone is a native speaker of English and *Experienced* equals 1 when someone has had previous tagging experience.

To examine if there is any interaction between the two variables, I calculated the ϕ correlation coefficient (Yule, 1912), which is a variation of Pearson’s product-moment correlation (ρ) aiming to establish the dependence between two binary categorical variables. For the following table, the ϕ coefficient is given by equation 5.1.

	<i>Native</i> = 0	<i>Native</i> = 1	Total <i>Native</i>
<i>Experienced</i> = 0	a	b	$a + b$
<i>Experienced</i> = 1	c	d	$c + d$
Total <i>Experienced</i>	$a + c$	$b + d$	

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} \quad (5.1)$$

The correlation coefficient between the two variables is $\phi = -0.008$, which clearly shows that the two are independent. This simplifies later measurements, which can be made with the variables considered separately and not in tandem.

5.6 Existing corpora

Below is a short review of existing corpora that were considered prior to the creation of the parallel corpus described in this chapter, along with the reasons why each one was dis-preferred.

A large-scale corpus that contains annotations of both tags and text is Social-ODP-2k9⁴ (Zubiaga et al., 2009). The corpus consists of 12,616 unique URLs (bookmarks) from Delicious, all of which are annotated with, among other things, the number of users tagging the bookmark, its top ten tags, user notes and reviews from StumbleUpon⁵. Despite its large coverage and tag-text parallel data, the corpus is not ideal for the task of investigating the existence of inter-tag relations, since **i**) Delicious notes do not amount to

⁴<http://nlp.uned.es/social-tagging/socialodp2k9/>

⁵<https://www.stumbleupon.com/>

descriptions; they simply constitute complementary information, **ii**) the resources (web-pages) typically contain text, which can influence the language of the submitted notes.

Two datasets from the Referring Expression Generation field, the GRE3D7 corpus (Viethen and Dale, 2011) and the Wally Referring Expression Corpus (WREC)⁶ (Clarke et al., 2013), contain textual descriptions of entities, used for modelling how humans select attributes that distinguish entities in an image from ‘distractors’. GRE3D7 is a collection of 4,480 descriptions of geometrical objects in different shapes, sizes, colours and positions. Such a dataset is too specific to be applied to folksonomy images, which have a variety of themes. WREC, contains 4,256 descriptions of individuals in 28 different real and visually complex scenes. It encodes wide-ranging relations between objects (e.g. ‘a man with lots of light blue hair and a grin’) but still lacks a tag-text duality, as it contains only text.

Finally, an interesting attempt at creating a tag and text parallel corpus is the work of Khan et al. (2012), from the field of video annotation. In their experiment, 140 videos from TREC data were annotated by 13 people with a title, keywords and a textual description (1,820 annotations in total), which could be used for improving video search. This corpus is the closest to the parallel corpus described in this chapter, with respect to its structure. In future experiments, it would be interesting to explore whether the interaction between tags and text in video annotation is similar to the one seen in image annotation.

5.7 Summary

In this chapter, I presented a parallel corpus of tags and text that I collected for a set of images through a human experiment. After discussing the experimental objectives, design and choice of stimuli, I described three pilot experiments. Then, I provided details of the final experiment, which was completed by 219 participants. Finally, I briefly reviewed some existing corpora, explaining why they were not suitable for this research.

⁶<http://datashare.is.ed.ac.uk/handle/10283/337>

Chapter 6

Implicit inter-tag relations

Having compiled a tags-text parallel corpus, we can now use the collected data to examine whether implicit semantic relations exist between tags and, if so, gain an insight into the nature of these relations. In Section 6.1, I examine the distribution of tags in participants' annotations and the distribution of words in the accompanying textual descriptions in order to ensure that they are similar to those observed in folksonomies and text corpora respectively, which will make any claims easier to generalise. In Section 6.2, I investigate the extent to which tags are found in descriptions of an image at both an individual (i.e. one's tags and their own descriptions) and a collective (i.e. all users' tags and all users' descriptions) level. In Section 6.3 I show how relations were extracted from the parallel corpus and discuss their nature. Finally, in Section 6.4, I outline some related work on inducing semantic relations from text and explain the decisions made with respect to extracting relations from the parallel corpus.

Data processing was restricted to a subcorpus which covers almost 70% of the entire parallel corpus, leaving the rest as potential test data for various implementations. This subcorpus was created after sampling 150 among the 218 participant IDs and recording all relevant details associated with them (e.g. tags, descriptions, tagging experience and so on). Sampling was performed in such a way that half of the participants in the subcorpus (i.e. 75) had performed tagging in Phase One and the other half had started the experiment with textual description. The reason for this was to ensure that the subcorpus is balanced, like the entire corpus, so that it renders potential order effects irrelevant for measurements. In the subcorpus, 70.6% of the participants are native speakers of English (106 out of 150) and 74% have no previous tagging experience (111 out of 150). These percentages are similar to those obtained for the entire corpus, which shows that the sample is representative of the complete data. This, in turn, indicates that results from any measurements performed for the subcorpus can be true of the entire corpus. As expected, the two participant variables from the subcorpus do not correlate ($\phi = 0.01$; see §5.5 for details of the correlation co-efficient).

6.1 Distributions of tags and words

Based on the discussion in Section 5.1, the parallel corpus – and, by extension, the subcorpus – can be treated as a folksonomy enriched with textual descriptions. The distribution of tags in this folksonomy can be examined, as was done for the Steve corpus in Section 3.1. Figure 6.1 reveals that tags submitted by all participants for all five images

are distributed in a way that approaches a zipfian distribution. This confirms the initial expectation that ‘consensus’ on the use of particular tags is reached even when users have no access to each other’s tags (see §3.1). A larger folksonomy would be expected to have a steeper decline of frequencies in the top ranks. A zipfian distribution can also be observed for words in description sentences (Figure 6.2), as well as tags assigned to a particular image by all participants in the subcorpus (Figure 6.3).

Tags and words were also counted with respect to images and participant variables. Unless specified, there are no significant differences between complementary participant groups (e.g. those with vs. those without tagging experience). Significance has been measured in all cases using the simulation-based significance test described in Section 4.2.2.

Tag tokens The average number of tag tokens per image per participant was 5.4 tags, 0.4 tags higher than the number of mandatory tag fields on the interface. If multi-word tags are split into individual tags by whitespace, the number slightly increases (5.5 tokens).

Word tokens The average number of word tokens in descriptions per image per participant is 53.1 words. Native speakers wrote longer descriptions (55.6 words on average) than non-native speakers (46.9 words). This difference is significant at a 99.2% confidence level.

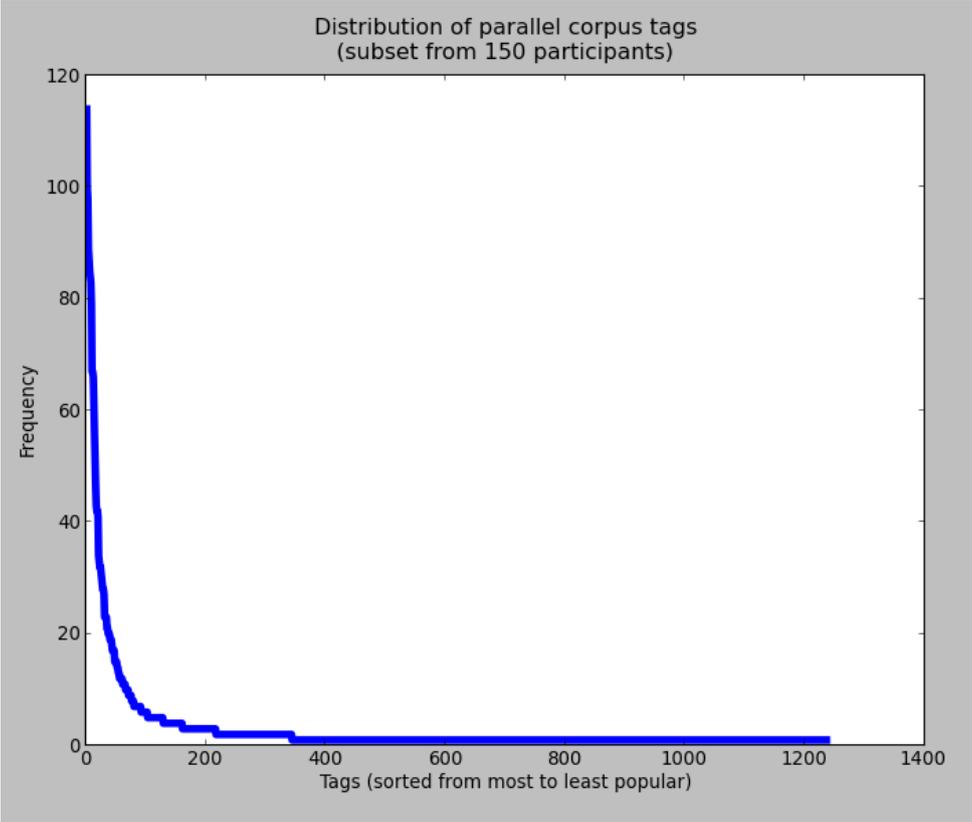


Figure 6.1: Distribution of tags from all five images

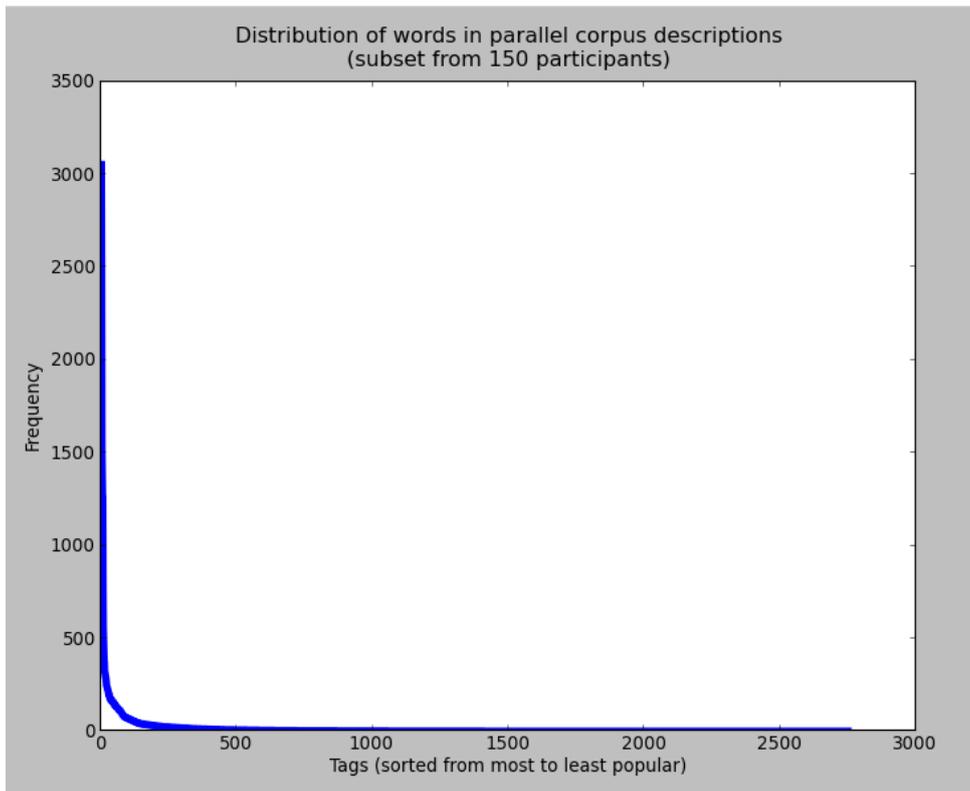


Figure 6.2: Distribution of description words from all five images

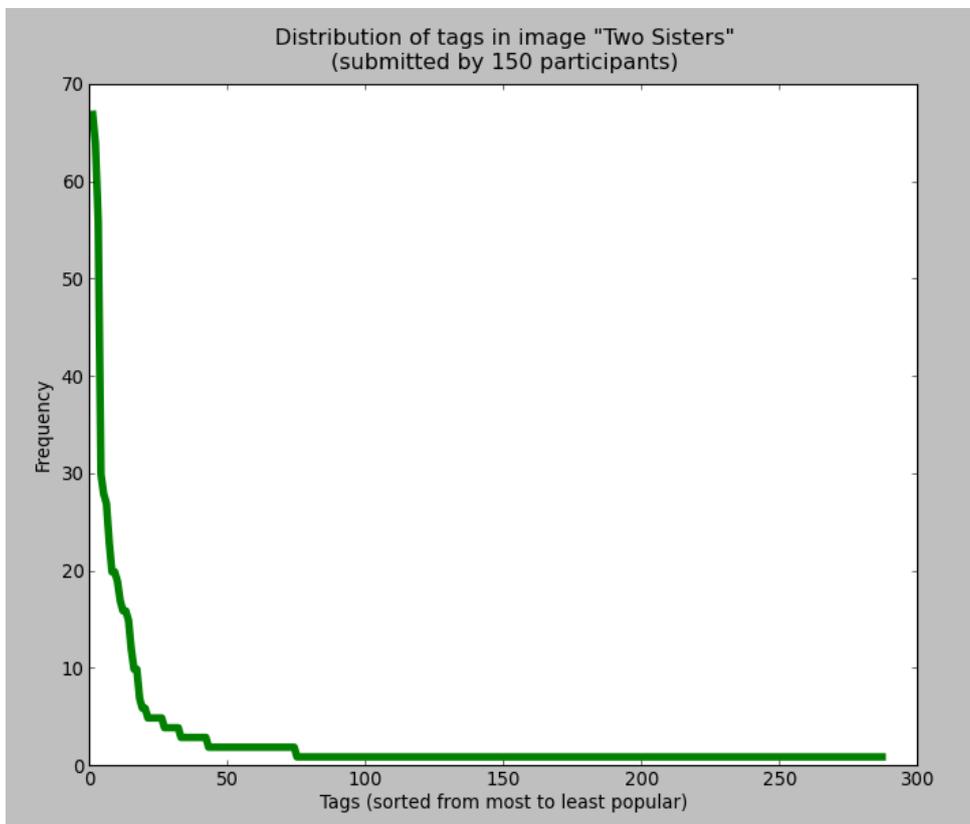


Figure 6.3: Distribution of tags in image "Two Sisters"

Tag hapaxes In the entire subcorpus, there were 893 tag hapax legomena out of 1,236 tag types (i.e. 72.2%). Spelling correction slightly reduces the number of hapaxes to 851 and the vocabulary to 1,195 tag types.¹ After spelling correction, the percentage of hapax legomena over the entire vocabulary reduced (71.2%), which shows that hapaxes were responsible for more mis-spellings than the rest of the types. Lemmatisation was also performed, using the Morpha lemmatiser (Minnen et al., 2000) for the inflectional morphology of English.² This step reduced the tokens to 788 and the types to 1,123, resulting in 70.1% tag hapaxes.

Word hapaxes The entire collection of descriptions in the subcorpus contains 1,330 word hapax legomena out of 2,721 word types (i.e. 48.9%). Spelling correction reduced the hapaxes to 1,188, the types to 2,578 and the percentage of the former over the latter to 46%. After lemmatisation, there were 908 hapaxes out of 2,098 word types (i.e. 43.2%). Some order effect was observed in this measurement: participants who completed the description task in Phase One had a significantly higher percentage of hapax words compared with those who started with tagging (37.5% vs. 31.1%), with a confidence level of 99.7%. Spelling correction increased the gap (35% vs. 28.6%) and therefore the confidence, which was now 99.9%. What this might suggest was that people who started the experiment with textual description tend to use more rare words, possibly because they take more care to produce precise descriptions. This order effect should not be considered a threat to the validity of any measurements, since the corpus is balanced, so, overall, there is no bias towards any particular tagging or describing behaviour.

6.2 Tag-word overlaps

Before attempting to explore whether and how tags relate with each other in text, it is necessary to answer some basic questions: To what extent do people use the same words to tag and to describe an image? To what extent do tag pairs associated with an image occur in the same sentences in the textual descriptions? These questions will be answered at both the *individual* and the *collective* level. The former involves comparing a user's tags for with her/his own descriptions, while the latter involves comparing the aggregated tags from all users for an image with all aggregated descriptions. Both levels are important for this research. The first level guarantees that the thoughts expressed in a given textual description belong to the person who produced the tags. The second level reflects the public opinion with respect to an image.

Single tag vs. single word overlap Before tags were processed, 43.7% of the tags used by a participant (*individual overlap*) for a given image were found in her/his description of the same image. An example parallel annotation from one participant can be seen in Figure 6.4, with overlapping tokens highlighted. After multi-word tags were split by whitespace into individual tags, the overlap increased to 53.6% and after lemmatisation

¹Spelling correction was performed using a variation of Peter Norvig's spelling corrector (<http://norvig.com/spell-correct.html>), which I trained on 6,708 wikipedia articles that were labelled with a category that contains the word "art" (case insensitive). The spelling correction module was evaluated against the Birkbeck spelling error corpus (Mitton, 1987). 83.4% of tokens that had been corrected by the module were found in the Birkbeck corpus with the same corrections.

²I used a Java implementation from <https://github.com/knowitall/morpha>

it reached 55.6%. This is evidence that people tend to use the same words to tag and describe an image. The overlap was also measured after dumping together all the tags for a given image (i.e. from all participants) and all the descriptions (*collective overlap*). With multi-word tags preserved, 53.2% of the tags used by all participants for each picture appear as words in the collected descriptions for the same image. The percentage rises to 73% after multi-word tags are separated and to 73.8% after lemmatisation.

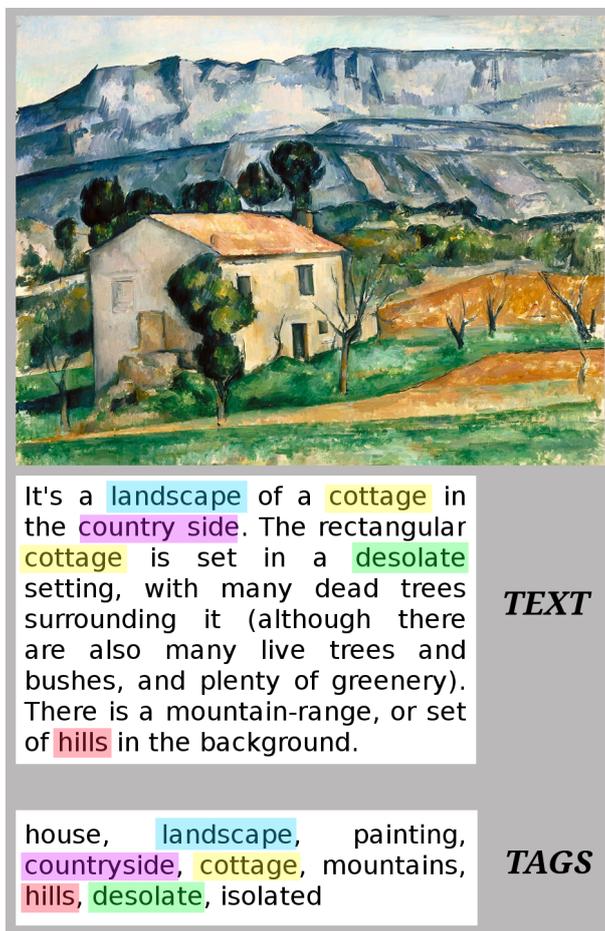


Figure 6.4: “House in Provence” and parallel data of Participant 33

Tag pair vs. word pair overlap Overlap between tag pairs in a participant’s annotation and word pairs inside sentences of same participant’s description was 10.9% initially, 17.5% after splitting multi-word tags and 18.8% after lemmatisation. Collective overlap was 15.9% initially, 35.9% after splitting multi-word tags and 39% after lemmatisation. These percentages are very high considering that, on average, a tag cloud for an image in this corpus contains 235 distinct tags, so it has $\binom{235}{2} = 27,495$ possible tag pairs.

The above result shows that co-occurrence of tags in tag clouds is sometimes the same as co-occurrence of words in sentences. If such co-occurring tags are linked with implicit semantic relations in the tag cloud, then it should be possible to use the text in which they also co-occur to examine whether such relations which were latent in the tagging have been made explicit in the descriptions.

6.3 Inter-tag relations in parallel corpus

Having shown that tag pairs of a given image in the corpus tend to be present in description sentences submitted for the same image, we can attempt to determine the extent to which co-occurring tags are connected with semantic relations in the supporting text, as well as to understand the nature of the relations that do exist.

To make explicit any implicit inter-tag relations, the obvious solution would be to perform relation extraction, with entities being tag pairs and relations being representations induced from the accompanying text. Another solution would be to crowd-source the relations by asking humans to select tag pairs and provide relations for them. However, instructing participants to provide relations that conform to a particular format (e.g. “Write at most three words between the two tags; a relation does not need to contain a verb” etc.) might involve time-consuming training, encourage participants to find more relations than those implicit in the tag cloud, while forcing them to perform an unnatural task. For this reason, performing relation extraction using the textual descriptions was the preferred approach.

Relation extraction for a pair of entities can be performed in two ways: **i)** with pre-determined types of relationships (e.g. one could look for entities that are connected with an *is-a* relationship, for interacting protein pairs etc.) or **ii)** with open-ended relationships (i.e. without prior knowledge of the types of relationships that may exist in the text). A review of related work in relation extraction can be found in Section 6.4.

As already mentioned, the nature of the relations that may exist between tag pairs is not known and should be explored through processing of the textual descriptions. For this reason, open-ended relation extraction is the preferred approach for inducing inter-tag relations from text.

In open-ended relation extraction, there are no relation-specific patterns or training data to ensure a satisfactory quality of induced relations, thus, deep syntactic processing might be needed. This is a view shared by many computational linguists working on similar problems. Wu and Weld (2010) argue that using parsed, as opposed to lightly processed, text in open information extraction improves performance. Lin and Pantel (2001) extract reliable relations by following dependency paths between two nouns of interest. Banko et al. (2007) also use dependency paths to induce relations that they consider ‘trustworthy’ and good enough to automatically train a classifier. *Dependency* representations in particular are useful for relation extraction since “certain semantic information is implicitly contained into dependency trees” (Herrera et al., 2006). A dependency tree (or graph) does not retain ‘horizontal’ (i.e. order) information about a sentence, as a phrase structure tree does, but it encodes ‘vertical’ (i.e. head-dependent) information³ that helps to “efficiently derive the core functor-argument structure of a sentence as an interface to semantic interpretation” (Hahn and Meurers, 2011). As Nirve (2005) explains, dependency structures “are less expressive than most constituency-based representations, but they compensate for this by providing a relatively direct encoding of predicate-argument structure, which is relevant for semantic interpretation”. To motivate the use of dependency parsing for relation extraction, Lin and Pantel (2001) explain that each arc between a head and a dependent amounts to a *direct* semantic relation between

³“the head identifies the meaningful object to which the meaning of a dependent contributes” (Kruijff, 2006)

two words, while a path consisting of arcs and nodes represents an *indirect* semantic relation between the two end words.

To discover inter-tag relations in the parallel corpus, I use dependency structures. One challenge is that tags do not belong to a fixed grammatical category, hence relation extraction techniques between nouns (e.g. (Lin and Pantel, 2001) and (Nakov, 2007)), noun phrases (e.g. (Banko et al., 2007) and (Etzioni et al., 2011)), named entities (e.g. (Hasegawa et al., 2004) and (Shinyama and Sekine, 2006)) cannot be used without modifications. To address this issue, I have specified a set of constraints that determine reliable semantic relations, namely, **i**) path constraints (§6.3.1) and **ii**) pattern constraints (§6.3.2).

The decisions taken in order to make explicit the latent inter-tag relations can be summarised as follows. First, relation extraction from textual descriptions was preferred over crowd-sourcing relations in the desired format, given that text **i**) provides an easier and more natural way for humans to express their thoughts on an image and **ii**) is likely to contain tag pairs connected with an explicit semantic relation. Second, extracting open-ended relations was preferred over extracting relations of pre-determined types, given that the types of semantic connections implicit between tag pairs were not known in advance. Finally, constraints were introduced in order to reduce the space of possible semantic relations, given that the entities (tags) for which relations are extracted are not restricted to a particular part of speech.

6.3.1 Path constraints

Below I describe how relations between tag pairs are discovered inside textual descriptions. To extract inter-tag relations from text, I utilised the grammatical paths that connect pairs of words in an image’s textual descriptions, when these words co-occur as tags in the same image’s tag cloud. I parsed all the descriptions of the 150-participant subcorpus (see beginning of this chapter) using Briscoe and Carroll-style dependencies (Briscoe and Carroll, 1995; Briscoe, 2006) via the C&C parser (Curran et al., 2007). Each sentence is a dependency graph and relations are assumed to be revealed by *paths* in this graph, which are labelled with Grammatical Relations (GRs) (Briscoe and Carroll, 1995). A dependency-parsed sentence can be seen in Figure 6.5, where the underlined words are tags.⁴

⁴Note that in tag-sequence grammars, which GRs (Briscoe, 2006) are native to, dependencies form a *graph* that allows for cycles, thus, not all sentences in a corpus will be parsed in a dependency *tree*.

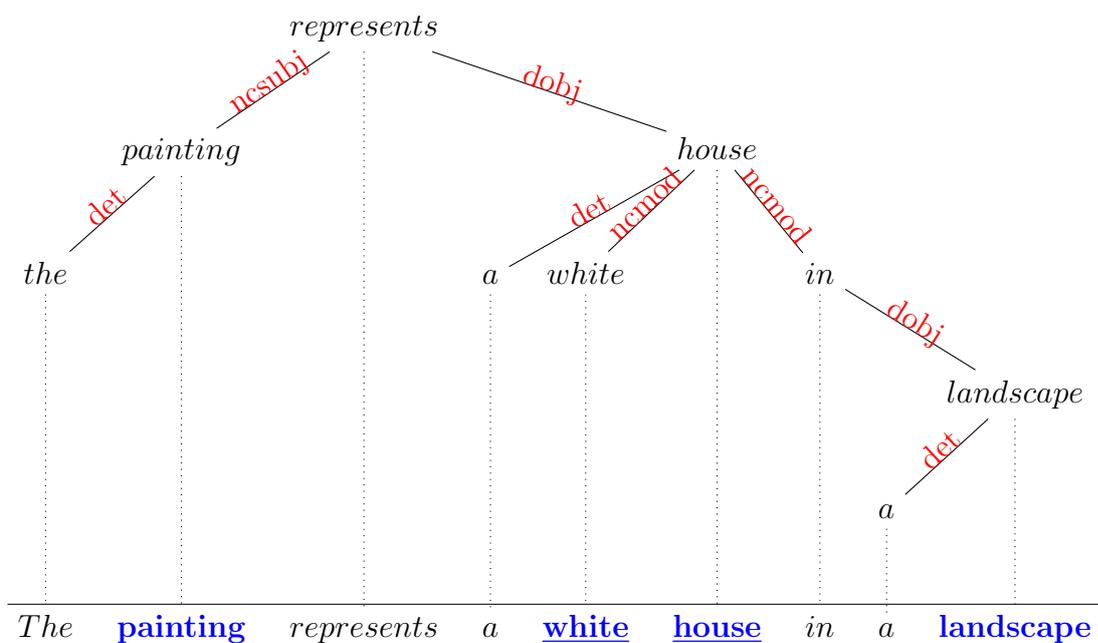


Figure 6.5: **Dependency tree example.** Words in blue are tags; ‘ncsubj’ stands for ‘non-clausal subject’, ‘det’ stands for ‘determiner’, ‘dobj’ stands for ‘direct object’ and ‘ncmod’ stands for ‘non-clausal modification’.

The paths of the dependency graph which will inform relation extraction from the parallel corpus data were selected according to the following rules:

1. A path should be *connected* and *without branching*.
2. *Both* end nodes of a path should be tags in the tag cloud of a particular image.
3. *Only* a path’s end nodes (i.e. no intermediate nodes) should be tags in the tag cloud of a particular image.
4. The *maximum* length of a path should be 4 nodes (3 arcs).
5. The *minimum* length of a path should be 2 nodes (1 arc).

Rule 1 specifies that a path should be a special case of a *catena* (i.e. ‘chain’, connected subgraph, in the dependency graph); in particular, a catena which has no branching (I call this ‘unbranched catena’). For instance, “white house in a landscape” (along with its arcs) is a legitimate catena but not a legitimate path for this task, since it does not form a single line. It would be unbranched if “a” had been excluded.

Rule 2 permits paths such as $house \xrightarrow{ncmod} in \xrightarrow{dobj} landscape$ but forbids paths such as $represents \xrightarrow{dobj} house \xrightarrow{ncmod} in \xrightarrow{dobj} landscape$, since ‘represents’ is not a tag.

Rule 3 excludes paths like $white \xleftarrow{ncmod} house \xrightarrow{ncmod} in \xrightarrow{dobj} landscape$ since an intermediate node (“house”) appears as a tag. If intermediate nodes were allowed to be

tags, then the extracted relations would have to be ternary or quaternary, which would require complex relation extraction rules. In addition, extracting multiple non-overlapping tuples (e.g. *white* \xleftarrow{ncmod} *house* AND *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *landscape*) involves only minimal information loss compared to the combined representation.

In **Rule 4**, the maximum path length attempts to minimise relations that are too specific, by disallowing the existence of more than two intermediate nodes, which could lead to paths such as *painting* \xleftarrow{ncsubj} *represents* \xrightarrow{dobj} *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *landscape*.

According to **Rule 5**, a two-node path can also have semantic value (e.g. *white* \xleftarrow{ncmod} *house*) based on the view that any head-dependent relationship encodes a semantic relation.

Applying the above restrictions, I extracted relations between tag pairs from the tag clouds of the five images in the parallel corpus. Instances of tag-relation-tag tuples, henceforth *Instantiated Dependency Patterns* (IDPs), can be of length 2 (i.e. two nodes; one arc), length 3 (i.e. three nodes; two arcs) or length 4 (i.e. four nodes; three arcs). As a result of the above path constraints, the leftmost and the rightmost node in each IDP are tags labelling the same image while the intervening nodes, if any, are not tags of the image in question. Sample IDPs with their frequency in image-specific text can be seen below.



length 2

<i>brush</i> \xleftarrow{ncmod} <i>strokes</i>	48
<i>naked</i> \xleftarrow{ncmod} <i>woman</i>	17

length 3

<i>painting</i> \xrightarrow{ncmod} <i>of</i> \xrightarrow{dobj} <i>woman</i>	8
<i>green</i> \xleftarrow{conj} <i>and</i> \xrightarrow{conj} <i>white</i>	4

length 4

<i>painting</i> \xleftarrow{ncsubj} <i>done</i> \xrightarrow{iobj} <i>with</i> \xrightarrow{dobj} <i>strokes</i>	2
<i>figure</i> \xrightarrow{cmod} <i>looks</i> \xrightarrow{iobj} <i>like</i> \xrightarrow{dobj} <i>woman</i>	1



length 2

blue \xleftarrow{ncmod} *mountains* 12
green \xleftarrow{ncmod} *trees* 12

length 3

painting \xrightarrow{ncmod} *of* \xrightarrow{dobj} *house* 20
house \xleftarrow{ncsubj} *is* \xrightarrow{xcomp} *white* 4

length 4

house \xleftarrow{ncsubj} *surrounded* \xrightarrow{ncmod} *by* \xrightarrow{dobj} *trees* 9
cottage \xrightarrow{ncmod} *along* \xrightarrow{iobj} *with* \xrightarrow{dobj} *mountains* 1



length 2

glass \xleftarrow{ncmod} *window* 12
white \xleftarrow{ncmod} *wings* 12

length 3

angel \xrightarrow{ncmod} *with* \xrightarrow{dobj} *wings* 11
window \xrightarrow{xmod} *featuring* \xrightarrow{dobj} *angel* 3

length 4

picture \xleftarrow{ncsubj} *is* \xrightarrow{iobj} *of* \xrightarrow{dobj} *window* 2
angel \xrightarrow{cmod} *dressed* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *armor* 1



length 2

<i>red</i>	\xleftarrow{ncmod}	<i>dresses</i>	65
<i>girls</i>	\xrightarrow{xmod}	<i>standing</i>	11

length 3

<i>girls</i>	\xleftarrow{ncsubj}	<i>wearing</i>	\xrightarrow{dobj}	<i>dresses</i>	22
<i>vase</i>	\xrightarrow{ncmod}	<i>of</i>	\xrightarrow{dobj}	<i>flowers</i>	21

length 4

<i>standing</i>	\xrightarrow{ncmod}	<i>next</i>	\xrightarrow{iobj}	<i>to</i>	\xrightarrow{dobj}	<i>table</i>	16
<i>table</i>	\xleftarrow{ncsubj}	<i>covered</i>	\xrightarrow{iobj}	<i>with</i>	\xrightarrow{dobj}	<i>tablecloth</i>	5



length 2

<i>moulin</i>	\xleftarrow{ncmod}	<i>rouge</i>	82
<i>woman</i>	\xleftarrow{ncsubj}	<i>dancing</i>	21

length 3

<i>man</i>	\xrightarrow{ncmod}	<i>in</i>	\xrightarrow{dobj}	<i>hat</i>	16
<i>men</i>	\xleftarrow{conj}	<i>and</i>	\xrightarrow{conj}	<i>women</i>	8

length 4

<i>people</i>	\xrightarrow{xmod}	<i>dressed</i>	\xrightarrow{ncmod}	<i>in</i>	\xrightarrow{dobj}	<i>hats</i>	2
<i>audience</i>	\xleftarrow{ncsubj}	<i>depicted</i>	\xrightarrow{iobj}	<i>as</i>	\xrightarrow{dobj}	<i>shadows</i>	1

6.3.2 Pattern constraints

To further control the quality of what counts as a tag-relation-tag instance, I introduced additional restrictions, called ‘pattern constraints’. We saw that each extracted path is an Instantiated Dependency Pattern. If these patterns are left with un-instantiated nodes (e.g. $* \xrightarrow{ncmod} * \xrightarrow{dobj} *$ instead of *painting* \xrightarrow{ncmod} *of* \xrightarrow{dobj} *house*), then they

provide more abstract information about the building blocks of inter-tag relations. I call these ‘Abstract Dependency Patterns’ (ADPs). Intuitively, ADPs can be instantiated into valid relation tuples if they are frequent enough; rare patterns could represent mistakes or uninteresting relations. Hence, I imposed a *pattern frequency constraint* whereby a tag-relation-tag instance (i.e. IDP) can only be considered valid if its ADP (i.e. its generalisation) occurs more often than a certain threshold in the text of the entire subcorpus. In other words, I regarded ADPs with a lower count as bad hosts for a tag-relation-tag instance. The threshold I set was 10.

An additional hypothesis is that, if an IDP can be fully decomposed into smaller valid IDPs, then it is less likely to capture the underlying relation between the two end tags. For example, an IDP such as *white* \xleftarrow{ncmod} *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *countryside* does not successfully capture a semantic relation between the tags “white” and “countryside”; this can be diagnosed by the fact that the IDP can be decomposed into smaller valid IDPs *white* \xleftarrow{ncmod} *house* and *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *countryside*. This observation can be generalised into abstract patterns (e.g. the ADP $* \xleftarrow{ncmod} * \xrightarrow{ncmod} * \xrightarrow{dobj} *$ is unlikely to host a good relation since it can be decomposed into the two smaller ADPs which are already valid: $* \xleftarrow{ncmod} *$ and $* \xrightarrow{ncmod} * \xrightarrow{dobj} *$). Hence, I imposed a *pattern redundancy constraint* whereby an ADP will only be considered a legitimate host for an inter-tag relation if it cannot be fully decomposed into ADPs that are more frequent than it is. For example, ADP $* \xleftarrow{ncmod} * \xleftarrow{conj} * \xrightarrow{conj} *$ occurs 37 times in the corpus but it can be replaced by two ADPs: $* \xleftarrow{ncmod} *$, which occurs 1,403 times and $* \xleftarrow{conj} * \xrightarrow{conj} *$, which occurs 230 times. Table 6.2 shows ADPs that occur more than 10 times, with patterns in bold being redundant, thus, eliminated.

Table 6.2: Top Abstract Dependency Patterns (ADPs), their counts in all descriptions and examples. ADPs in bold-face have been filtered out.

PATTERN	COUNT	EXAMPLE
$* \xleftarrow{ncmod} *$	1403	blue mountains
$* \xrightarrow{ncmod} * \xrightarrow{dobj} *$	779	vase with flowers
$* \xleftarrow{conj} * \xrightarrow{conj} *$	230	angel and trumpet
$* \xleftarrow{ncsubj} * \xrightarrow{dobj} *$	150	picture showing cottage
$* \xleftarrow{ncsubj} * \xrightarrow{ncmod} * \xrightarrow{dobj} *$	113	house surrounded by trees
$* \xleftarrow{ncsubj} * \xrightarrow{xcomp} *$	108	panels are bright
$* \xleftarrow{ncsubj} *$	101	lady dancing
$* \xleftarrow{ncmod} * \xrightarrow{ncmod} * \xrightarrow{dobj} *$	79	blue tower in background
$* \xleftarrow{ncsubj} * \xrightarrow{iobj} * \xrightarrow{dobj} *$	61	picture looks like advertisement
$* \xrightarrow{xmod} * \xrightarrow{dobj} *$	60	vase with flowers
$* \xleftarrow{ncmod} * \xleftarrow{conj} * \xrightarrow{conj} *$	37	religious motifs and colours

* $\frac{iobj}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	33	sisters in dresses
* $\frac{cmod}{\rightarrow}$ * $\frac{xcomp}{\rightarrow}$ *	28	lady is naked
* $\frac{ncmod}{\rightarrow}$ * $\frac{iobj}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	26	cottage along with mountains
* $\frac{xmod}{\rightarrow}$ * $\frac{ncmod}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	26	sky painted on glass
* $\frac{dobj}{\rightarrow}$ *	24	dancing cancan
* $\frac{xmod}{\rightarrow}$ * $\frac{iobj}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	16	woman painted with strokes
* $\frac{xmod}{\rightarrow}$ * $\frac{xcomp}{\rightarrow}$ *	14	figure is abstract
* $\frac{ncmod}{\leftarrow}$ * $\frac{ncsubj}{\leftarrow}$ * $\frac{dobj}{\rightarrow}$ *	13	impressionist style depicting person
* $\frac{cmod}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	12	cathedral depicts angel
* $\frac{ncmod}{\rightarrow}$ *	12	standing nude
* $\frac{ncmod}{\leftarrow}$ * $\frac{ncsubj}{\leftarrow}$ *	12	primary colours yellow
* $\frac{ncsubj}{\leftarrow}$ * $\frac{xmod}{\rightarrow}$ * $\frac{dobj}{\rightarrow}$ *	12	woman depicted using reds

6.3.3 Overlaps and compositionality

Overlaps

In Section 6.2 we saw that, without lemmatisation or splitting multi-word tags, 15.9% of all possible tag *pairs* in an image co-occur as two words in descriptions of the same image. Now that it is clear what counts as a semantically related tag pair on the basis of path and pattern criteria, it is possible to gain an understanding of how many of the overlapping tag pairs are indeed related in sentences. Using the approved ADPs, I extracted relations between tag pairs that occur in sentences. I found that approximately 1/3 of the overlapping tag pairs (i.e. 4.9% of all possible tag pairs) were connected with semantic relations.

Compositionality

This chapter has provided evidence that semantic relationships may hold between particular pairs of tags. In other words, some tag pairs behave compositionally. However, it is not known whether this composition happens because of underlying syntactic restrictions between the tags or despite the absence of such restrictions.

To detect possible traces of underlying syntax between the tags, I examined the order in which individual users submit their tags for a given image. In particular, I compared the flexibility of *tag submission order* within pairs of tags from all users' annotations of an image with the flexibility of *word order* within the corresponding pairs of words in all users' descriptions of the same image. I only considered tag pairs that were connected with valid semantic relations, as specified in the previous sections. A first observation was

that, even in cases where the order of words in a two-word phrase from the descriptions is essentially *fixed* (e.g. words “young” and “girls” appeared only as “young girls” in textual descriptions and not as “girls young”), when tags are elicited, they may be given in *either order* with equal or similar frequency (e.g. “girls”-“young” 50%; “young”-“girls” 50%). Another example could be “vase” and “flowers” appearing as words in descriptions (“vase”-“flowers” 93%; “flowers”-“vase” 7%, in phrases such as “vase of flowers”, “vase full of white flowers”, “vase with flowers”) and as tags (“vase”-“flowers” 64%; “flowers”-“vase” 36%).

An analysis of the entire tag and description datasets revealed that the order between two tags co-occurring in a users’ annotations was at least 5 times more flexible than the order of the same pairs appearing as semantically connected words within text descriptions. Tag (or word) order flexibility is defined as $P(ab) - P(ba)$, where $P(ab)$ is the probability of sequence a-b (e.g. 64% for “vase”-“flowers”) and $P(ba)$ is the probability of the reverse sequence b-a (e.g. 36% for “flowers”-“vase”). The difference between tag order flexibility and word order flexibility is significant at a higher than 99.99% confidence level.

Such a result might suggest that tags are largely unordered (or, more plausibly, that tag order is largely irrelevant to users), thus it is possible that the underlying composition observed between them occurs despite the absence of syntactic restrictions. In other words, it can be said that *semantic* relations are not necessarily *syntactically* constrained, which might point in the direction of dynamic compositionality, as understood within the contextualist school of thought in Linguistics (Travis, 1997, 2000; Recanati, 2004); a compositionality not strictly and not only bound by syntax.

6.4 Related work

Below is some previous work on extracting relations between entities in text, parts of which influenced the relation extraction method described in this chapter.

6.4.1 Pre-specified relations

Early work in relation extraction aimed at inducing relations from text via manually written patterns, tailored to a particular pre-determined relation of interest. For example, Hearst (1992) extracted instances of the *hyponymy* relation from text (i.e. X-Y pairs whereby X is a hypernym of Y) using patterns like “X such as Y” and “such X as Y”. Berland and Charniak (1999) extracted examples of the *meronymy* relation (i.e. X-Y pairs whereby X is part of Y) with patterns like “X of a Y” and “Y’s X”. Hand-crafting extraction patterns was labour-intensive and suffered from low coverage, since there were many more patterns for a relation type than a human could pre-specify.

A solution to manually writing relation extraction patterns was to use bootstrapping, a semi-supervised method whereby a small set of known instances of a relation (e.g. “Marie Curie” – “radium” and “Christopher Columbus” – “America” as good examples of the “*discovered*” relation) are fed into the system to help it learn the relevant patterns from text (e.g. “discovery of Y by X”), which can then be used to extract more instances of the relation. Brin (1999) used this method to extract patterns for the “*is author of*” relation by providing the system with seed author-work tuples like “Charles Dickens” – “Great Expectations” and “William Shakespeare” – “The Comedy of Errors”.

Another method to extract relations from text was through supervised learning, in which training text is annotated with pre-specified relation types (Zhou et al., 2005; Kim and Moldovan, 1993; Riloff, 1996; Soderland, 1999). A relation extraction system can learn a classifier that decides which relations are exemplified by stretches of unseen text on the basis of features that are either hand-written, or induced by kernels (Zelenko et al., 2003). However, producing training data is a costly procedure and, like bootstrapping, it requires that examples of each relation type are provided to the system.

To eliminate the need for manually annotating documents with relations, Craven and Kumlien (1999) utilise “weakly labelled” training data by pairing documents with entity-relation-entity triples that they are likely to assert. The researchers collect subcellular localisation facts for proteins (i.e. which part of a cell a protein is located in; e.g. “collagen” – “is located in” – “extracellular matrix”) from the Yeast Protein Database (YPD) (Hodges et al., 1998). A subset of those facts include references to abstracts of academic articles which mention the given subcellular localisation information, while a subset of those abstracts mention both entities in question. This reduced set of abstracts, along with their corresponding YPD facts was a weakly supervised corpus that they used to train a classifier. Mintz et al. (2009) suggest “distant supervision”s, whereby one can create training data automatically by matching entity-relation-entity triples from large ontologies (e.g. Freebase⁵) with sentences that contain both entities. Although this method is bound to introduce noise, it can create a very large training corpus without any human involvement.

6.4.2 Open-ended relations

A different line of work in relation extraction is focused on inducing relations from text without determining their types in advance. This process, known as relation discovery, is ideal in cases where deciding what relations are present in the text is difficult or *ad hoc*. For instance, extracting relations between tag pairs, as described in this chapter, had to be open-ended, since there was no way of knowing in advance what types of relations to expect.

Lin and Pantel (2001) performed open-ended relation extraction by harvesting ‘X-path-Y’ triples from text, where X and Y represent the entities in question and path is a chain of grammatical *dependencies* obtained from the MINIPAR parser (Lin, 1993) and intermediate *words* as in $X \xleftarrow{N:subj:V} find \xrightarrow{V:obj:N} solution \xrightarrow{N:to:N} Y$ which is paraphrasable as “X finds solution to Y”. Acceptable entity-relation-entity triples are the X-path-Y chains that satisfy the following constraints: **i)** X and Y are nouns, **ii)** dependencies should only connect content words (i.e. nouns, verbs, adjectives and adverbs), **iii)** only paths that occur more often than a certain threshold are kept. The latter restriction eliminates the need to restrict path lengths, since long paths are unlikely to exceed the threshold. The authors then go on to cluster the paths using what they call the ‘extended distributional hypothesis’, that is, “If two paths tend to occur in similar contexts, the meanings of the paths tend to be similar”. In other words, they create what Lin had called ‘thesauri’ in (Lin, 1998) (see §4.3.2), extended to relation paths. Relations like “X resolves Y”, “X finds solution to Y” and “Y is solved through X” are considered similar (i.e. ‘synonyms’ in the same thesaurus entry) with path similarity being judged on the basis of features such as ‘has w1 in Slot X’ or ‘has w2 in Slot Y’.

⁵<https://www.freebase.com/>

Hasegawa, Sekine and Grishman (2004) introduce what has been called ‘on-demand information extraction’. The authors cluster pairs of named entities (NE) of pre-approved kinds (e.g. COMPANY – COMPANY or PERSON – GPE⁶) according to their distributional similarity. Similarity of an NE pair is based on cosine distance of vectors whose features are words between the two named entities in text. Term weighting is based on tf-idf (Salton et al., 1975) and similar NE pairs are considered those which belong to the same cluster after hierarchical clustering. Each cluster is assumed to contain instances (NEs connected with text such as “is President of”, “has served as the Mayor of”, “was the first female to be elected president of”) of a discovered relation type. Words that appear between most NE pairs of the same cluster were regarded as words in common; among these words, the most frequent was automatically selected as the label of the relation (e.g. “president”). A variant of this research was Shinyama and Sekine’s (2006) ‘unrestricted relation discovery’ whereby clusters are created for NE pairs without any restrictions on combinations of named entity types.

Banko et al. (2007) introduced ‘Open Information Extraction’ with the TextRunner system, which extracts open-ended relations from web-scale corpora. The system has three components:

1. *Self-supervised learner*, which extracts entity-relation-entity tuples from thousands of sentences parsed with a ‘deep linguistic parser’ (Klein and Manning, 2003) and automatically labels them as ‘trustworthy’ if they fulfil the following constraints: **i)** the two entities, which are extracted as ‘bases noun phrases’ (i.e. ones that have no nested noun phrases or modifiers), should be connected in a dependency graph with a maximum length; **ii)** these dependency paths cannot extend across two clauses and **iii)** none of the two entities can be pronouns. Tuples that fail to meet the above conditions are labelled as ‘untrustworthy’. The automatically labelled data are used to train a Naïve Bayes classifier with lightweight features such as POS tags, number of tokens in the relation string and the presence or absence of stopwords.
2. *Single-pass extractor*, which extracts entity-relation-entity triples from a 9 million webpage corpus by tagging each word with its most likely part of speech and performing shallow chunking to identify entities (e.g. ‘the Eiffel Tower’). Relations are found from the text intervening the two entities with the use of heuristics that eliminate ‘non-essential’ modifiers such as adjectives and prepositional phrases. When triples are extracted, they are presented to the previously trained classifier, which retains only reliable ones.
3. *Redundancy-based assessor*, which eliminates tuples that occur below a certain threshold in the entire corpus.

Etzioni et al. (2011) introduce ReVerb, a variant of TextRunner which imposes additional constraints for open information extraction in order to eliminate relations that are *incoherent* (e.g. “contains omits” from the sentence “The guide contains dead links and omits sites”) or *uninformative* (“took” instead of “took place in”). Syntactic constraints improved precision by requiring that extracted relations match particular pre-approved POS-tag patterns.⁷ Lexical constraints impose the rejection of relations that have a below-

⁶GPE stands for ‘Geo-political entity’.

⁷Despite the use of hand-written patterns, this is still open-ended relation extraction since patterns are not relation-specific.

threshold number of possible arguments. ReVerb achieves a 30% larger area under curve (AUC) in the recall-precision graph compared to TextRunner.

Open-ended relation extraction has also been attempted for noun-noun compounds (e.g. ‘tear gas’). A notable example is work by Nakov (2007), who induces paraphrase-type relations for nouns of interest by searching Google with wildcard queries such as ‘N2 that * N1’ (e.g. “gas that brings tears”, “gas that produces tears”), where N1 and N2 can be inflectional variants (e.g. plurals) of the first and second noun in the compound respectively and “that” can be replaced by “which” or “who”. One asterisk retrieved one word but queries were submitted with up to eight asterisks.

6.4.3 Relation extraction in folksonomies

The relation extraction techniques presented above, especially those that performed open-ended relation discovery, have influenced the methodology I used in this chapter. However, none of these techniques has been designed with folksonomy tags in mind.

In the folksonomy literature, research has concentrated on extracting hierarchical relations (subsumption, instantiation and equivalence), that are assumed to exist between folksonomy tags *across* resources. Hierarchical relations are typically induced with graph-based techniques (Heymann and Garcia-Molina, 2006; Benz and Hotho, 2007) or Association Rule Mining (Schmitz et al., 2006; Lin et al., 2009) and attempt to reveal a taxonomy which is believed to have emerged from a folksonomy.

A small number of studies have been conducted on extracting open-ended (as opposed to pre-defined, usually taxonomical) relations between folksonomy tags, as a way to reveal not just a taxonomy but a fully fledged ontology which is assumed to underlie the folksonomy (Specia and Motta, 2007; Maala et al., 2007; Angeletou et al., 2008; Sordo et al., 2010). However, these works have used some type of ontology as a corpus, a process which suffers from data sparsity since structured data, in contrast to tagging data, are hard to acquire. More importantly, none of them has attempted to extract relations with respect to particular resources, which is central to this thesis.

A general-purpose text corpus has been used for a small part of the research by Trabelshi et al. (2010) to help extract inter-tag relations across resources. The authors provide the details of an algorithm, which starts by discovering which concepts are *related*: two or more tags are seen as related if they are used by the same users on the same resources (i.e. they are part of a ‘tri-concept’). The components of the system are: **i)** pre-processing (consolidating tags, grouping similar tags, grouping synonym tags and filtering infrequent tagsets), **ii)** grouping related tags (by mining frequent tri-concepts) and **iii)** extracting relations from Wikipedia sentences parsed into phrase structure trees; each pair of related tags is checked against noun phrases while relations are extracted from verb phrases. The process was applied to a Delicious dataset, providing generally intuitive results. Evaluation was difficult because of the non-availability of gold standards. After performing some qualitative analysis, the authors noticed that some triples are nonsensical and that some relations seemed to be synonymous.

The folksonomy literature was not easily applicable to the problem of extracting image-specific inter-tag relations from the parallel corpus, as seen in this chapter. Relation discovery within NLP provided a better basis for such a task.

6.5 Summary

In this chapter, I analysed the parallel corpus of tags and text in order to determine whether implicit relations hold between tag pairs. First, I observed that both tags and words follow a power-law distribution. Then, I showed that a high proportion of tags (in isolation but also in pairs) appear in sentences of the textual descriptions. This legitimised the use of the text as a resource for harvesting inter-tag relations. I defined a semantic relation as a dependency path that connects a pair of tags in the text, given some constraints. Approximately one third of the tag pairs found in sentences were connected to each other with such relations. Open-ended relation extraction was used, but other methods were also reviewed.

Chapter 7

Postulating tag-relation-tag triples

In the previous chapter, I provided evidence for the existence of implicit relations between tag pairs, after processing textual descriptions that had been submitted by participants along with their tags for each one of five images. Such text is a more detailed record of users' thoughts with regard to an image than tags are. This allows for the recovery of semantically connected tag pairs and their implicit relations to be uncovered. However, in real-life tagging systems, text rarely accompanies tags and, if it does, it tends to be in the form of notes and not fully-fledged descriptions. In the absence of textual data, tag-relation-tag triples can only be postulated. In this chapter, I explore ways in which 'candidate' triples can be induced for a given image using text corpora which do not describe the image in question. In Section 7.1, I describe the task and discuss its feasibility. In Section 7.2, I outline the constraints placed on a text corpus for the purpose of extracting high-quality triples. Finally, in Section 7.3, I provide an account of my attempts to improve the quality of proposed tag-relation-tag triples by using specialised corpora, as opposed to a general-purpose corpus. Evaluation of the final system will be described in Chapter 8.

7.1 Task description

7.1.1 Overview

This chapter discusses the process of building a *proof-of-concept* system, whose aim is to demonstrate that it is possible to: **i)** detect pairs of semantically connected tags from an image's tag cloud and **ii)** suggest relations between them, in the absence of text describing the specific image.

The first goal, that is discovering tag pairs for which relations hold with respect to an image, is important because, rather than showing simple co-occurrence patterns (see Chapter 4), it exposes patterns of semantic connectivity. Even before specifying the nature of the relation, identifying the right tag pairs can provide valuable information. For example, Figure 7.1 (henceforth "Coney Island") shows an image and its tag cloud. If this tag cloud is converted to a graph of semantically connected tag pairs, where node-arc-node paths are equivalent to tag-relation-tag triples (see Figure 7.2), it can provide an alternative, more human interpretable, display of the original tags. Figure 7.2 is a labelled directed graph, where node size represents the number of times a tag has been used as an endpoint in a tag-relation-tag triple generated for the image. A graph representation

can be more informative than a tag cloud. For instance, it can show which tags would be vital as words if the image were to be described in full text. In the original tag cloud, “beach” is more important than “people” because, presumably, the image is more about a beach than it is about people. Yet, the opposite is true in the graph, because many of the tags annotating the image are semantically connected with “people” but few are connected with “beach”. Tags that are central in the graph may be seen as ‘theme’ (that is important elements for the description of the image) while more peripheral ones may be seen as ‘rheme’ (see discussion in §2.3.2).

Figure 7.1: “Coney Island” (1934) by Paul Cadmus



20thcentury acrobats acrobatsbeachgoers actionpacked
 atmosphericperspective atthebeach balloon bathingsuits **beach**
 beachscene beer caos carnival carnivale cartoon classicism colourful comum
 paulcadmus coneyislandamerican confusion congestedrecreation
crowdcrowdedfigures crowded rollercoaster
 figures flesh frantic gayart gluttony grossrepresentation grotesque groupportraitinlandscape
 humanpyramid humor icecreamcone lazy leisure lowlife luridstyle magicrealism
 marsh masses newyork oil oiloncanvas orgy outdoor painting
peopleredblloon popular prominentnoses pyramid redballoon
 rollercoasterbackground sand satire seabackground sexuality shore skybackground
 beachgoers summerfun sunburn survivalofthestrone swine takingaphoto thirties
 tower **toweroiloncanvasoil** unitedstates
 cyclonerollercoaster vibrant washingtonbaths 1934 **people**
 crowd **coneyisland** debauchery drunken excess
 allages american stereotypes summer hitler holdingforsupport hot
 urbanscene manyassortedposes manybodytypes
unitedstatesbeach

7.1.2 Feasibility

In order to assess the feasibility of postulating high-quality tag-relation-tag triples without image-specific text or any access to visual information, I performed a small-scale experiment with one participant. The specific objective was to decide whether it is possible for a human to guess *which* tags relate with each other and what the *nature* of their relation is, without consulting the image.

The participant was shown two tag clouds, one after the other, without the associated images. For the first tag cloud (Figure 7.3), the participant was asked to write some textual description of the hidden image, based on the tags. It was explained that: **i)** tag size represents how many users annotated the image with a given tag and **ii)** the task is to provide text that is as truthful as possible with respect to the image.

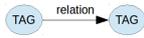
20thcentury arms beautiful bliss blue **breeze** calypso caminando classical **cloud** clouds
contemplation dress ethereal european fieldofflowers floressalvajes **flower** flowers georgehitchcock **grass** greek
greekmythology green hair hill **hillside** innocent lady life montana mujer natural nuves odyssey oiloncanvas parrish
partlycloudy paysaje **peaceful** profile purpleflowers purpleribbon **ribbon rock** rockoutcropping rocks
royalblue seducedbyzeus skin sky spring summer sunshine tarde traditional tragicstory victorian viento virginal vrouw **white**
whitedress wildflower wildflowers windy **woman** womanonhill wreath

Figure 7.3: Tag cloud for painting “Calypso” (1906) by George Hitchcock

The participant’s response was: “A young woman wearing a white dress with a purple ribbon (wreath?) on her hair. She’s walking on a rocky hillside. She looks like a character from Greek mythology. Painting from late XIX or early XXth.” Figure 7.4 shows the image, which verifies that the description is successful to a good extent.



Figure 7.4: Painting “Calypso”

For the second tag cloud (Figure 7.5), the participant was asked to draw on paper some tag-relation-tag triples in the following form: . As an example, I provided the triple **woman** $\xrightarrow{\text{wearing}}$ **hat**, from a hypothetical image. I omitted dependency patterns from the example triple in an attempt to avoid complication. The participant had the option of providing a set of separate triples or creating a network of connected triples. No instructions were given as to tag synonymy, polysemy or multi-word tags.

amusing anarchist arrogant assertive baroque **black** blackhair blackshirt
 caravaggesque cassero castle chair clergy cleric clever count court deliberate
 forceful french frenchpainter **gesture** gesturing goatee hands highlights humor
 male maleportrait **man** mazarin megalomania menicucci monte montesansavino
paper personality pope **portrait** pretentious priest proud
 roccadelconte roman san savino scholarly **seated** serious squinting staring teacher
 moustache mustach mustache oil vanity vigorous white wiseman words wrinkles
 boulogne buffoon buffoonery buttons canvas
 questioning reasoning religiousboss ego expatriate expression eyebrows fame
 tired unabashed urbanviii vain valentin immediacy italian **jester** largehands light

Figure 7.5: Tag cloud for painting “Rafaello Menicucci” (1625) by Valentin de Boulogne

The participant produced a network, which can be seen in Figure 7.6. Again, the response verifies that most of the triples are relevant to the image (Figure 7.7), which was revealed afterwards.

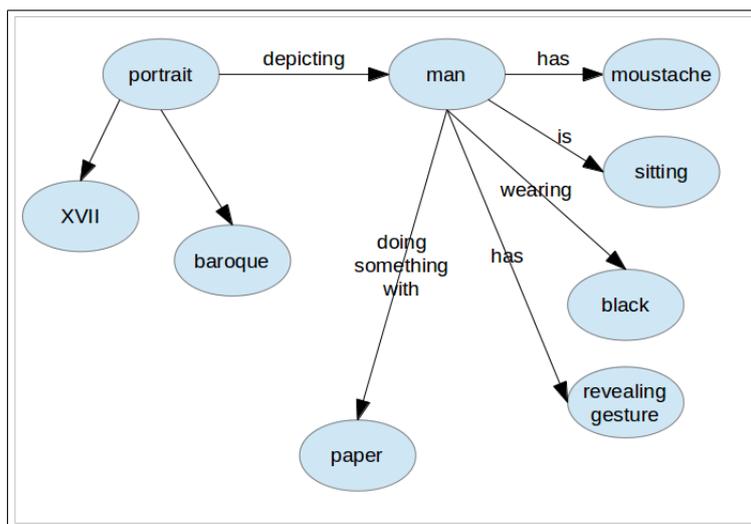


Figure 7.6: Network of tag-relation-tag triples for painting “Rafaello Menicucci”



Figure 7.7: Painting “Rafaello Menicucci”

This experiment suggested that tags seen together in a tag cloud provide more information to a human than the individual meanings of the tags. This, in turn, indicates that it might be possible to estimate the implicit relations between tags without access to the image itself.

7.2 Producing triples

In this Section, I describe two attempts towards the design of a system which aims to generate tag-relation-tag triples without relying on image-specific text. In this research, triples have been treated as equivalent to Instantiated Dependency Patterns (IDPs), as explained in Section 6.3.1. For a generated IDP to be considered of high-quality, it should fulfil two requirements: well-formedness and plausibility.

An IDP is *well-formed* if it is syntactically correct and semantically sound. A syntactically correct IDP **i)** is extracted from a sentence which a native speaker of English would judge as grammatical and **ii)** has not resulted from a parser error. A semantically sound IDP is the analogue of a semantically sound sentence, that is one which does not display “oddness” (Cruse, 1986). Oddness can be caused by features such as pleonasm (i.e. redundancy; e.g. “a female mother”), dissonance (i.e. flouting of selectional preferences; e.g. “Kate was very married.”), improbability (i.e. a statement being unlikely to be true in most situations; e.g. “The kitty drank a bottle of claret”) or zeugma (i.e. a word being used with two senses simultaneously; e.g. “Arthur and his driving licence expired last Thursday”) (Cruse, 1986). Examples of well-formed IDPs are $white \xleftarrow{ncmod} house$ and $city \xrightarrow{ncmod} in \xrightarrow{dobj} 1950s$, which are both syntactically correct and semantically sound. Examples of ill-formed IDPs are:

- $people \xleftarrow{ncsubj} is \xrightarrow{xcomp} popular$ (extracted from the ungrammatical Wikipedia sentence “People from the island is popular for shark fishing and working in cargo vessels”)
- $tree \xrightarrow{ncmod} to \xrightarrow{dobj} tall$ (resulting from a mis-parsing; originally from the sentence “It is a medium-sized tree to tall”)
- $green \xleftarrow{ncsubj} born \xrightarrow{ncmod} in \xrightarrow{dobj} california$ (improbable, hence semantically odd)¹

¹extracted from the sentence “Melody Green was born in Hollywood, California on August 25 1946”.

An IDP is *plausible* with respect to an image if a human can judge it as true of this image. For instance, *people* \xrightarrow{ncmod} *on* \xrightarrow{dobj} *beach* and *american* \xleftarrow{ncmod} *painting* are plausible with respect to the image “Coney Island” (p. 98) because it is highly likely that humans could judge them as true. On the other hand, the IDP *people* \xleftarrow{ncsubj} *visited* \xrightarrow{dobj} *tower*, despite being well-formed, is implausible for this image.

Below I describe two systems, ‘Dep’ and ‘POS-Dep’, and explain why the latter was chosen as more suitable for the task. Both systems extract IDPs from text (with end nodes being tags in an image’s tag cloud). To ensure that IDPs are of high quality, the two systems apply different constraints with respect to well-formedness and plausibility. In the rest of the thesis, I will be using the term *acceptable* to refer to IDPs that are both well-formed and plausible with respect to a given image.

7.2.1 Well-formedness constraints

The first system created, called ‘Dep’ (for ‘dependency’), produces triples in the form of IDPs for images in the Steve corpus, using Wikipedia. To guarantee well-formedness, the system requires that an IDP should be extracted from text only if it can unify with an Abstract Dependency Pattern (ADP; see §6.3.2) from a pre-approved set. The process is as follows:

- From the Steve corpus, only images which contain more than 30 distinct tags were kept. Among the total of 33,948 images in the corpus, only 1,496 contained enough tags. For the purposes of this chapter, I will be using examples from a sample of three images: “Coney Island” (Figure 7.1, p. 98), “Detroit” and “Grizzly Giant Sequoia” (both on Figure 7.8, p. 105).
- Some tags were considered stopwords and were removed from the tag clouds. This was a quick and easy way to avoid function words (such as prepositions) as end nodes for tag-relation-tag triples. Although such words are rarely used as tags, extracting IDPs with them as end nodes can overload the system without offering in return any informative or well-formed relation. The stopword list was constructed manually and consists of 92 items (including articles, prepositions, pronouns, single letters, auxiliary verbs etc).²
- Multi-word tags were not used as potential end points for tag-relation-tag triples, but were useful for a later stage (see §7.2.2)
- The set of ADPs learnt from the parallel corpus (p. 90) was used to extract all tag-relation-tag triples (IDPs) from Wikipedia (October 2013 version), which contains approximately 2.7 billion tokens.³ Wikipedia was parsed with the C&C parser (Curran et al., 2007) outputting Grammatical Relations (Briscoe and Carroll, 1995), the same output used while extracting relations from the parallel corpus (§6.3.1).
- IDPs that occur less than 5 times in the entire Wikipedia were excluded. The intuition was that an extracted IDP which occurs very infrequently is likely to be

²The complete list can be found on <http://www.cl.cam.ac.uk/~tt309/exper/stopWords>

³The text was downloaded from <http://web.archive.org/web/20131027044816/http://dumps.wikimedia.org/enwiki/20131001> and the corpus statistics are from <http://en.wikipedia.org/wiki/Wikipedia:Size.comparisons>

syntactically incorrect (e.g. arising from grammatical or parser errors), or semantically odd, therefore unlikely to capture useful semantics. For instance, the IDP $distance \xrightarrow{ncmod} between \xrightarrow{dobj} oil$, postulated for the “Grizzly Giant Sequoia” (Figure 7.8b) occurs once in Wikipedia and is not a good example of a potential semantic relation between “distance” and “oil”, even before considering the particular image for which it is extracted. This IDP is extracted from the sentence “The water puts physical distance between the oil or pan and the proteins on the surface of the meat”. Another hapax-mentioned IDP suggested for the same image is $tall \xleftarrow{ncmod} nature$, from the sentence “Each and every participating speaker is given three to five minutes to give a short speech of a tall tale nature, and is then judged according to several factors”.

- From all the above-threshold IDPs extracted, the system kept the ones whose end tags co-occur in tag clouds of the images for which relations were sought. It, then, produced an initial ranking for every image, sorting the corresponding IDPs from the most to the least frequent in the corpus. The count of an IDP in the corpus initially appears to be a good predictor of well-formedness.

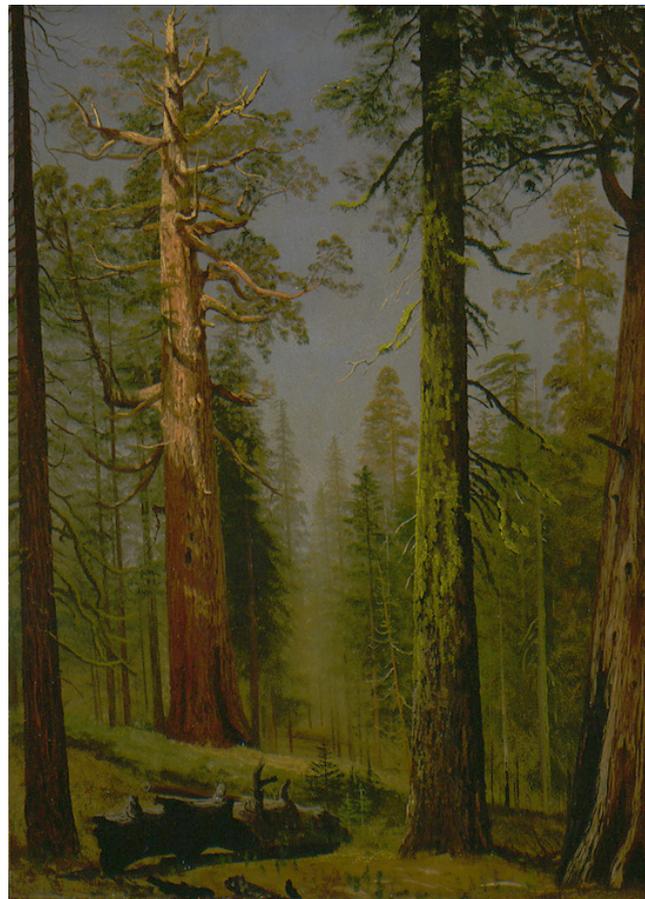
The number of above-threshold IDPs produced by the Dep system varies for each image depending on the size of the tag cloud and the frequency of its tags when they appear as words in Wikipedia. The Dep system suggested 257 IDPs for “Coney Island”, 430 IDPs for “The Grizzly Giant Sequoia” and 1,606 IDPs for “Detroit”. The top IDPs for “Grizzly Giant Sequoia”, sorted by their count in Wikipedia, are: $oil \xleftarrow{ncmod} painting$ (occurs 1,926 times), $dark \xleftarrow{ncmod} green$ (1,435 times), $landscape \xleftarrow{ncmod} painting$ (958 times) and $tall \xleftarrow{ncmod} trees$ (685 times). In “Coney Island”, the top IDPs are $american \xleftarrow{ncmod} people$ (2,227 times), $oil \xleftarrow{ncmod} painting$ (1,926 times), $crowd \xrightarrow{ncmod} of \xrightarrow{dobj} people$ (1,395 times) and $hot \xleftarrow{ncmod} summer$ (612 times). For “Detroit”, the top IDPs are $black \xleftarrow{conj} and \xrightarrow{conj} white$ (6,813 times), $white \xleftarrow{conj} , \xrightarrow{conj} black$ (6,508 times), $old \xleftarrow{ncmod} city$ (3,618 times), $1940s \xleftarrow{conj} and \xrightarrow{conj} 1950s$ (2,664 times) and $old \xleftarrow{ncmod} buildings$ (2,291 times).

Ill-formed triples Although the output of the Dep system tends to be well-formed, there are also a number of ill-formed IDPs produced, typical examples of which can be seen in Table 7.1. The first five examples in the table were suggested by the system for “Coney Island”, the next two for “Detroit” and the rest for “Grizzly Giant Sequoia”.

An interesting case is the IDP $photograph \xrightarrow{dobj} people$ postulated for the image “Detroit”. This IDP is found 35 times in Wikipedia, which suggests that it is unlikely to be a parser error. Indeed, as shown in the example sentence, the IDP is perfectly well-formed if “photograph” is treated as a verb. However, the image had been annotated with the tag “photograph”, presumably because it is a photograph, rather than, say, a painting; hence, the tag acts like a noun. Even without knowledge of the image that the tag is associated with, verbs are known to be unusual in a tag cloud (see Section 3.2). Therefore, when “photograph” is a noun, which is typically the case in a tag cloud, the IDP $photograph \xrightarrow{dobj} people$ is ill-formed. Such an observation leads to the conclusion that well-formedness should be judged *with respect to* part-of-speech tags.



(a) "Detroit, 1943" by Harry Callahan



(b) "The Grizzly Giant Sequoia, Mariposa Grove, California" (1872) by Albert Bierstadt

Figure 7.8: Sample images from Steve corpus; selected randomly among those labelled with at least 30 distinct tags

Table 7.1: Examples of ill-formed IDPs suggested by the Dep system, their count in Wikipedia and example source sentences

ILL-FORMED IDP	COUNT	SAMPLE SENTENCE
$american \xleftarrow{ncmod} hot$	17	It also made it onto the <i>American Billboard Hot 100</i> .
$people \xleftarrow{conj} , \xrightarrow{conj} figures$	22	Counties are most often named for <i>people</i> , often political <i>figures</i> .
$crowd \xrightarrow{ncmod} - \xrightarrow{dobj} people$	9	A reported <i>crowd</i> of 100 – 150,000 <i>people</i> attended.
$american \xleftarrow{conj} and \xrightarrow{conj} people$	14	The Sioux are Native <i>American</i> and First Nations <i>people</i> in North America.
$summer \xleftarrow{ncsubj} be \xrightarrow{xcomp} hot$	9	<i>Summer</i> can be uncomfortably <i>hot</i> and humid.
$city \xrightarrow{ncmod} : \xrightarrow{dobj} street$	16	Art galleries are springing up on many streets across the <i>City</i> : <i>James Street</i> , <i>King William Street</i> [...] to name a few.
$photograph \xrightarrow{dobj} people$	35	Rarely did he <i>photograph people</i> or make portraits.
$tree \xrightarrow{ncmod} to \xrightarrow{dobj} tall$	8	It is a medium-sized <i>tree to tall</i> .
$wood \xleftarrow{ncsubj} mounted$	11	Perforated <i>wood</i> may be <i>mounted</i> as a thin strip.

As seen above, the Dep system attempts to ensure well-formedness by requiring that IDPs conform to the Abstract Dependency Patterns (ADPs) learnt from the parallel corpus. For instance, $crowd \xrightarrow{ncmod} of \xrightarrow{dobj} people$ is allowed because it can unify with the ADP $* \xleftarrow{ncmod} * \xrightarrow{dobj} *$, learnt from text describing tagged images. However, such ADPs are too permissive if used in a general-purpose corpus like Wikipedia. For instance, $* \xleftarrow{ncsubj} * \xrightarrow{xcomp} *$ allows IDPs such as $summer \xleftarrow{ncsubj} be \xrightarrow{xcomp} hot$ (Table 7.1), whose underlying POS-tags for “summer”, “be” and “hot” in the example sentence are “NN” (singular common noun), “VB” (verb in base form) and “JJ” (adjective) respectively.⁴ Such IDPs may have been eliminated if the pre-approved Abstract Dependency Patterns were less abstract, hence less permissive. The ADPs used so far have specified edges (e.g. \xleftarrow{ncmod}) but unspecified nodes (e.g. in $* \xleftarrow{ncmod} *$). However, ADPs can be learnt with more information in mind. For instance, nodes can be partially specified, having POS tags as constraints (e.g. $NN \xleftarrow{ncsubj} VBZ \xrightarrow{xcomp} JJ$), which can make unification with an IDP more difficult. The nodes of an IDP have the POS tags that the original words had in the Wikipedia sentence. These POS tags need to match those of a more restrictive ADP for the IDP to be accepted.

⁴The tagset used is from the Penn Treebank (<https://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html>), as output from the C&C part-of-speech tagger, prior to super-tagging with CCG (Combinatory Categorical Grammar) lexical categories and parsing (Curran et al., 2007).

To learn POS-tag-informed dependency patterns from the parallel corpus, I used the same ‘path constraints’ and ‘pattern constraints’ as the ones described in Section 6.3, however, instead of simple place-holders for dependency path nodes, I extracted the original POS tags from the sentences, which will impose further constraints on what grammatical constructions an IDP can be extracted from in the future. The ranking of the patterns was done with the POS tags as part of the pattern (e.g. $JJ \xleftarrow{ncmod} NN$ and $JJ \xleftarrow{ncmod} NNS$ were considered two different patterns). POS-tag-informed ADPs that were found less than five times in the text of the parallel subcorpus were eliminated, as were those that did not fulfil ‘pattern redundancy constraints’ (§6.3.2). All the above-threshold ADPs can be seen in Table 7.2.

The POS-tag-informed patterns learnt from the textual descriptions of the parallel subcorpus were used to extract IDPs from Wikipedia with a system called ‘POS-Dep’. The POS-Dep system follows the exact same process as Dep but imposes slightly more sophisticated constraints on dependency patterns without significantly increasing the processing cost, given that dependency-parsed corpora are usually already POS-tagged.

When run on the Steve corpus, POS-Dep produces fewer IDPs than Dep, as expected. In each one of the three example images, the POS-Dep system prevents almost half of the IDPs produced by the Dep system, eliminating many ill-formed triples (e.g. $american \xleftarrow{ncmod} hot, crowd \xrightarrow{ncmod} - \xrightarrow{dobj} people$ etc.), while not sacrificing many of the well-formed ones. For this reason, POS-Dep was the only system used for subsequent experiments in this chapter. Evaluation of this system is described in Chapter 8.

Table 7.2: POS-tag-informed dependency patterns extracted from the parallel subcorpus. Patterns in boldface were later eliminated by automatically applying a ‘pattern redundancy constraint’ (§6.3.2).

PATTERN	COUNT	EXAMPLE
$JJ \xleftarrow{ncmod} NN$	269	white house, blue mountain
$NN \xrightarrow{ncmod} IN \xrightarrow{dobj} NN$	176	view of cottage, window from church
$NN \xleftarrow{ncmod} NN$	129	glass artwork, oil painting
$JJ \xleftarrow{ncmod} NNS$	119	rough strokes, red dresses
$NN \xrightarrow{ncmod} IN \xrightarrow{dobj} NNS$	75	painting of girls, house with trees
$CD \xleftarrow{ncmod} NNS$	62	two girls
$JJ \xleftarrow{conj} CC \xrightarrow{conj} JJ$	48	blue and yellow, abstract but realistic
$NN \xleftarrow{ncsubj} VBG$	47	audience standing, girl dancing
$NN \xleftarrow{ncmod} NNS$	43	brush strokes, glass windows

$NN \xleftarrow{conj} CC \xrightarrow{conj} NN$	40	grass and hill, house or cottage
$NN \xleftarrow{ncsubj} VBG \xrightarrow{dobj} NN$	35	man wearing hat, angel holding trumpet
$NNS \xrightarrow{ncmod} IN \xrightarrow{dobj} NN$	34	shadows of audience, mountains in background
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{xcomp} JJ$	29	background is colourful, painting seems cold
$VBG \xrightarrow{ncmod} IN \xrightarrow{dobj} NN$	28	standing beside table, dancing on stage
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{dobj} NN$	21	figure has eyes, painting depicts torso
$NNS \xrightarrow{ncmod} IN \xrightarrow{dobj} NNS$	20	girls in dresses, mountains behind fields
$NNP \xleftarrow{ncmod} NN$	20	moulin poster, red paint
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{xcomp} NN$	20	dancer is dress, mountain is house [<i>artefact of parsing "There is" sen- tences</i>]
$NNS \xleftarrow{ncsubj} VBG$	18	girls standing
$NNP \xleftarrow{ncmod} NNP$	18	moulin rouge
$NN \xrightarrow{ncmod} IN \xrightarrow{dobj} NNP$	17	club in paris, painting by cézanne
$NNS \xleftarrow{ncsubj} VBP \xrightarrow{xcomp} JJ$	17	fields are yellow, girls appear forelorn
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{iobj} IN \xrightarrow{dobj} NN$	16	picture looks like poster, figure is of angel
$JJ \xleftarrow{ncmod} NN \xrightarrow{ncmod} IN \xrightarrow{dobj} NN$	16	religious depiction with an- gel, african lady in tribe
$NNS \xleftarrow{conj} CC \xrightarrow{conj} NNS$	13	fields and trees, strokes and colours
$NNS \xleftarrow{conj} CC \xrightarrow{conj} NN$	11	mountains and house, flowers and chair
$NN \xleftarrow{ncsubj} VBN \xrightarrow{ncmod} IN \xrightarrow{dobj} NNS$	11	house surrounded by fields, figure overlaid by strokes
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{dobj} NNS$	10	window has words, painting uses strokes
$NN \xleftarrow{conj} CC \xrightarrow{conj} NNS$	9	grass and hills, landscape and fields

$NN \xleftarrow{ncmod} NN \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NN	9	brass pot with plant, glass window in church
$NN \xleftarrow{ncsubj} VBN \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NNS	9	vase placed on table, scene surrounded by nature
$JJ \xleftarrow{ncmod} NN \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NNS	8	red top with dots, brush size with colours
$NNS \xleftarrow{ncsubj} VBG \xrightarrow{dobj} NNS$	8	girls wearing dresses, girls wearing gowns
$NN \xleftarrow{ncsubj} VBN \xrightarrow{iobj} IN \xrightarrow{dobj} NN$	8	table adorned with cloth, angel dressed as knight
$NNS \xleftarrow{ncsubj} VBP \xrightarrow{xcomp} NN$	7	colours are yellow, fields are green
$NN \xrightarrow{cmod} VBZ \xrightarrow{dobj} NN$	6	angel holds trumpet, poster shows dancing
$VBG \xrightarrow{ncmod} JJ \xrightarrow{iobj} TO \xrightarrow{dobj}$ NN	6	standing next to table
$NN \xleftarrow{ncsubj} VBZ \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NN	6	vase stands on table
$VBG \xrightarrow{ncmod} IN \xrightarrow{dobj} NNS$	5	standing on rocks, dancing in hats
$JJ \xleftarrow{ncmod} NN \xleftarrow{conj} CC \xrightarrow{conj}$ NN	5	spanish villa and sky
$NNS \xleftarrow{ncsubj} VBN \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NN	5	girls dressed in dress
$NNS \xleftarrow{ncsubj} VBN \xrightarrow{ncmod} IN \xrightarrow{dobj}$ NNS	5	fields surrounded by trees

7.2.2 Plausibility constraints

Postulating inter-tag relations that are plausible for a particular image in the absence of image-specific text or visual information is especially challenging. However, an attempt can be made to decide on the plausibility of each IDP based on heuristics (e.g. ‘IDPs whose distributional vectors are similar to the tag cloud are likely to be relevant’ or ‘IDPs that are very frequent in the corpus are likely to be true of most images’). Below I describe the criteria I have used to rank IDPs according to plausibility. These criteria, namely **i)** corpus count, **ii)** distributional similarity, **iii)** tag weight, **iv)** tag stems and **v)** multi-word tags, are later combined in an equation.

Corpus counts Intuitively, tag-relation-tag triples extracted from a text corpus will be true of a particular image if there is some degree of predictability. For example, in most contexts, tags “house” and “countryside” are related with an “in” relation, as in *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *countryside*. Predictable IDPs represent facts that are generally, or usually, true, so they are ones with a high count in a given corpus. More frequent IDPs are more likely to be relevant (all other things being equal) because a frequent IDP

reveals a common way in which particular words (or tags) relate to each other, which is also likely to manifest in an image.

Distributional similarity As another measure, a system could use Distributional Semantics to eliminate implausible IDPs. The assumption is that IDPs extracted from contexts which are more similar to the context provided by a tag cloud are more likely to be relevant to an image. In other words, an IDP is descriptive of a particular image when the contexts in which it tends to occur (i.e. its corpus-based distributional vector) is similar to the specific context in which the corresponding tags occur (i.e. the rest of the tag cloud, if seen as a vector). More details on this criterion will be provided at the end of this section.

Tag weight Another criterion for deciding whether an IDP is plausible with respect to an image is tag weight, which is a function of the number of times each one of the two end tags have been used in the image. The assumption is that tags used by more users on the same image are more likely to be connected with implicit relations and are more likely to be good descriptors of the image.

Tag stems Lemmatising the end nodes (tags) of an IDP can provide information on the the IDP’s plausibility with respect to an image. For instance, IDPs whose end tags have the same morphological base form (e.g. *trees* \xleftarrow{ncsubj} *include* \xrightarrow{dobj} *tree*) are less likely to be plausible. Expressions such as “trees include tree” are a side-effect of dependency paths ‘jumping over’ words that are usually important for an informative reading (e.g. “these *trees* include a palm tree”).

Multi-word tags Although multi-word tags (MWTs) were not used as end-points for IDPs (see §7.2.1), they can be useful for deciding on an IDP’s plausibility. MWTs often contain phrases, some of which are parseable as tag-relation-tag triples similar to those produced by the POS-Dep system. When the nodes of such IDPs exist in MWTs in the same order, they provide a clear indication that an IDP is plausible. For example, tags “blackandwhitephotography” (Figure 7.8a), “oilpapermountedboard” (Figure 7.8b) and “crowdcrowdedfigures” (Figure 7.1) correctly predict the plausibility of the extracted IDPs *black* \xleftarrow{conj} *and* \xrightarrow{conj} *white*, *mounted* \xleftarrow{ncmod} *board* and *crowded* \xleftarrow{ncmod} *figures* respectively. The above criteria were used to compute a plausibility score for each IDP. The score was given by the function $f(c, s, t, b, m)$, where:

- c is the corpus probability of the IDP (i.e. count in corpus divided by the added counts of all phrases)
- s is the distributional similarity between the vector constructed for the IDP from the corpus ($0 \leq s \leq 1$) and the tag cloud, treated as a vector.
- t is the ‘tag weight’, equivalent to the count of the first tag (leftmost node of IDP) plus the count of the second tag (rightmost node of IDP) divided by the number of tag tokens in the tag cloud.

- b is the ‘base form weight’, which is equivalent to 0 (i.e. penalisation) when the two end tags of the IDP in question have the same morphological base form and 1 when they have a different base form.
- m is the ‘multi-word tag weight’, which is 1 when the nodes of the IDP occur within a MWT in the same order and 0 otherwise.

The function was expressed as a linear combination of the above five variables, all of which had values from 0 to 1:

$$f(c, s, t, b, m) = a_1c + a_2s + a_3t + a_4b + a_5m \quad (7.1)$$

where a_1 , a_2 , a_3 , a_4 and a_5 are parameters (coefficients).

After experimentation with different images, I manually assigned values to the above parameters, which were used to produce a plausibility-based, as opposed to a purely count-based, ranking of postulated IDPs. In a later part of the thesis, I will show how these parameters can be optimised.

Distributional similarity as a predictor of plausibility

As mentioned above, we can attempt to predict the plausibility of an IDP for a particular image through distributional similarity. The tag cloud, being a multiset, can be treated as a word vector that acts as a surrogate for the meaning of the image. This vector can be compared to a feature vector that can be learnt for the IDP whose plausibility is measured. If the two vectors are similar enough, then the IDP is considered plausible for the image.

A feature vector for an IDP can be constructed from the corpus sentences in which the IDP occurs. This assumes that an IDP has one sense in the corpus. For instance, even if “oil” and “company” are ambiguous words, the triple $oil \xleftarrow{ncmod} company$ is not: when “oil” collocates with “company” it tends to mean fuel as opposed to, say, olive oil and when “company” collocates with “oil” it tends to mean business, and not being accompanied by people. This assumption is compatible with Yarowsky’s (1993) idea of “one sense per collocation”, used for word sense disambiguation. Since the two words are unambiguous when occurring together, the IDP itself is expected to have one meaning across different sentences.

With the POS-Dep system, distributional vectors were constructed for all IDPs that satisfied well-formedness constraints. The vectors used Wikipedia words as features and co-occurrence in sentences as values. An IDP co-occurs with a word if this word is found in the sentence but is not a node of the IDP or a stopword.⁵ For instance, in the sentence “The oil company was no longer on strike”, the IDP $oil \xleftarrow{ncmod} company$ co-occurs with the words “longer” and “strike”.

Frequently occurring IDPs produced vectors large enough to reliably decide whether the IDP is similar enough to the tag cloud (i.e. plausible for the image). However, for less frequent IDPs, vectors were sparse. This – combined with the fact that the tag clouds themselves typically consist of less than 100 distinct tags – renders vector similarity hard to obtain with confidence. To improve the coverage of distributional vectors, I used a

⁵The stopword list is the same as the one mentioned earlier in this chapter.

simplified version of the method described in (Melamud et al., 2013). In the paper, Melamud et al. extend lexical vectors by adding, for every word (feature) w_i in the vector, its N most similar words, where N is a fixed number, independent of the weight of w_i . Then, these N words, along with w_i (i.e. $N + 1$ words) share the original weight of w_i . For example, if $N = 5$, then for a word like “coffee” with weight 2.3 in the original vector, the top five most similar words are added (e.g. “cup”, “drink” etc.) and the original weight of 2.3 is divided equally between all the $5 + 1$ words. The method I used to expand tag clouds was simpler: for every tag in a tag cloud, I added as many top-similar words as the count of the tag. For instance, if the tag “flower” has been used by 15 people (i.e. it has count 15 in the tag cloud), then I added to the tag cloud the 15 most similar words to “flower” (e.g. “plant”, “leaf” etc). A tag with a smaller count would have fewer similar words added for it. This preserves the original weight of the tags while adding more words to the distribution. The similar words were taken from the distributional similarity-based thesaurus built from Wikiwoods, which I described in Section 4.3.2. The same method was used to expand sparse feature vectors for IDPs.

Using the above method, I computed the distributional similarity of each IDP extracted by the POS-Dep system with that of the tag cloud of the image in question. In the example images used in this chapter, distributional similarity was particularly effective at detecting the non-plausibility (i.e. low similarity with the tag cloud) of IDPs such as:

- (Figure 7.1) $sand \xleftarrow{ncmod} people$ (which occurred in Wikipedia as “Sand People”, referring to characters from the movie series *Star Wars*)
- (Figure 7.8b) $paper \xleftarrow{ncsubj} published \xrightarrow{iobj} in \xrightarrow{dobj} nature$ (which occurred in sentences referring to an academic paper published in the journal *Nature*)
- (Figure 7.8b) $tree \xleftarrow{ncmod} oil$ (from sentences referring to the essential oil “tea tree oil”)
- (Figure 7.8b) $paper \xleftarrow{ncmod} board$ (from sentences about a board of directors in a newspaper)

One obvious question regarding distributional vectors for IDPs is ‘why not use *compositional* distributional semantics?’ (Mitchell and Lapata, 2010), since compositionality accounts for the productivity of language and does not suffer from sparsity problems. For instance, a good representation of the IDP $oil \xleftarrow{ncmod} company$ could be constructed with a multiplicative model between the individual vectors of “oil” and “company”, which would eliminate or penalise words like “olive” from the vector of “oil” (since it rarely occurs in the context of “company”) or words like “together” from the vector of “company” (since it does not tend to occur in the context of “oil”). However, compositional distributional semantics would not be able to construct the right vector for cases like “Sand People” above, since the part of the meaning which refers to the movie series *Star Wars* does not arise from the composition of “sand” and “people”. Semantic meaning is not always compositional (e.g. in the case of idioms) or is sometimes weakly compositional (e.g. in the case of fixed expressions). Constructing feature vectors for entire IDPs is able to capture such non-compositional effects.

7.3 Using specialised text corpora

So far, only a general-purpose corpus, Wikipedia, has been used in order to harvest potential tag-relation-tag triples using dependency patterns. In this Section, I examine whether the POS-Dep system can suggest more acceptable triples through the use of specialised corpora, in particular an image caption corpus (§7.3.1) or a domain (visual arts) corpus (§7.3.2).

7.3.1 Image caption corpus

To extract acceptable IDPs, the system requires a corpus which is either large, such as Wikipedia, or dense with respect to the vocabulary of interest, such as the textual descriptions of the parallel corpus (see Chapter 5). Currently available image caption corpora are generally small, but likely to be more ‘tag-dense’ than general-purpose corpora. Since image captions typically contain words referring to objects, events or moods in an image, they are also likely to contain a large number of words that have been used as tags for images.

One widely used image caption corpus is the Pascal Sentence dataset (Farhadi et al., 2010), which consists of 1,000 images matched with five human descriptions each, that the researchers collected using Amazon’s Mechanical Turk. While this dataset is clean and potentially tag-dense, it is too small for extracting full IDPs. For instance, the tags “house” and “countryside” co-occur 8 times in the parallel sub-corpus (Figure 5.1a, p. 64) and 39 times in Wikipedia, while in Pascal they never co-occur. In fact, the individual word “countryside” is observed only five times.

A larger dataset is Im2Text, comprising one million image-caption pairs. Ordonez et al. (2011) assembled this corpus by crawling Flickr⁶ for image captions and automatically eliminating the ones that are less *visually* associated with their corresponding images. For example, the caption “This is a toddler playing with a duck” is visually descriptive, while a caption such as “Another good hobby for children” does not explicitly describe the objects or events in the image. Im2Text is, to my knowledge, the biggest image caption corpus reported in the literature. Kuznetsova et al. (2013) produce a cleaner version of Im2Text, called “Generalised 1M image caption corpus” (henceforth G1M)⁷, whose captions contain only visually salient information, omitting unnecessary details from sentences.

G1M was chosen as the most appropriate image caption corpus for the extraction of IDPs, since it is the most similar to the parallel corpus text, in that it prioritises visual description (e.g. describing what the image shows to someone who cannot see it; see §5.4.2). This corpus is also large enough to allow for extraction of some IDPs. In G1M (as well as in Im2Text) captions can consist of more than one sentence. Hence, before extracting the triples, I performed some simple sentence segmentation, using the regular expression `\\.\\s[A-Z]` (i.e. period, then whitespace, then capital letter) as a predictor of a sentence boundary. Then I parsed each sentence using the C & C parser, outputting Grammatical Relations (as in Sections 7.2.1 and 6.3.1). The POS-Dep system used this parsed corpus to suggest tag-relation-tag triples, discussed later in this chapter.

⁶<https://www.flickr.com/>

⁷Both Im2Text and Generalised 1M image caption corpus are available from <http://www.cs.stonybrook.edu/~pkuznetsova/imgcaption/>

7.3.2 Visual arts corpus

Creating a visual arts corpus can allow extraction of triples that are more likely to be accurate descriptors of arts images than those found in a general-purpose corpus. For example, for an image described with tags “stroke”, “patient” and “brush”, one would expect such a domain corpus to favour IDPs like *brush* \xleftarrow{ncmod} *stroke* and dis-prefer IDPs like *patient* \xrightarrow{ncmod} *with* \xrightarrow{dobj} *stroke*. The ideal visual arts corpus would be one with a higher density in art words. One consideration is the trade-off between density and coverage; smaller domain corpora might be more dense but have low coverage and vice-versa.

To my knowledge, there is no Visual Arts corpus available, hence, I decided to create one by extracting a subcorpus of Wikipedia which contains visual arts articles. I used the October 2013 Wikipedia dump; in particular, *text* data and *hierarchy* data.⁸ In order to create the subcorpus, I isolated articles annotated with categories which fall under the Wikipedia category ‘Visual Arts’. Then I extracted their text using the Wikipedia Extractor.⁹

Wikipedia articles fall under 28 types (namespaces).¹⁰ For the purpose of creating the domain corpus, the ‘main’ namespace (i.e. typical articles with texts; called ‘pages’) and the ‘category’ namespace (i.e. articles of the type ‘category’) were used.

Categorisation in Wikipedia is similar to collaborative tagging: users use uncontrolled category tags (e.g. ‘Arts’) to annotate typical articles (pages), implying that a given article is *about* a given category. In addition, Wikipedia category tags can annotate category articles, thus creating a category hierarchy. The resulting hierarchy is not a tree but a directed graph which allows for cycles and multiple ‘parent’ categories.¹¹

Given that the notion of article categorisation is loose, traversing the graph from a root category will give similar categories only in the first few steps; the further away one moves from the root category, the less relevant the categories are. For example, one path of depth 5 is *Visual arts* \rightarrow *Art materials* \rightarrow *Bronze* \rightarrow *Bronze Age* \rightarrow *Babylonia* \rightarrow *Akkadian language*. Since the relation between a category and its super-categories is not *is-a* but something equivalent to ‘falls under the topic of’, it is not surprising that ‘Akkadian language’ ends up being in a sub-graph of the visual arts category graph. For this reason, I traversed the category graph breadth-first and stopped at depths 2, 3, 4 and 5, creating four alternative visual arts corpora. To avoid following cycles in the graph, I prevented any category that has been visited during graph traversal from being visited again.

Once I extracted the relevant sub-categories, I collected articles that are labelled with them, for inclusion in the domain corpus. The four candidate domain corpora created were:

- “DepthTwo”
- “DepthThree”

⁸The 2013 Wikipedia dump is available on <http://web.archive.org/web/20131027044816/http://dumps.wikimedia.org/enwiki/20131001>. The *text* data and *hierarchy* data used are downloadable as [enwiki-20131001-pages-articles.xml.bz2](#) and [enwiki-20131001-categorylinks.sql.gz](#) respectively.

⁹http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

¹⁰<http://en.wikipedia.org/wiki/Wikipedia:Namespace>

¹¹<http://en.wikipedia.org/wiki/Wikipedia:Categorization>

- “DepthFour”
- “DepthFive”

Each one of the above corpora is a collection of articles that are tagged with the ‘eligible’ subcategories of the ‘Visual Arts’ category. Eligible categories are the ones that are not administrative (e.g. “1911 Britannica articles needing updates from March 2011”) and occur within a given depth from the root category node in the category hierarchy graph. For instance, for the creation of DepthTwo, the only articles considered were the ones tagged with categories that occur up to and including depth 2 from “Visual Arts” (root), depth 1 being the subcategories of “Visual arts” and depth 2 being their respective subcategories. Likewise, the categories chosen for the creation of the DepthThree corpus are the root category and the subcategories from the next three levels.

The four corpora have the following sizes:

- DepthTwo: 0.08% of tokens in Wikipedia
- DepthThree: 1.05% of tokens in Wikipedia
- DepthFour: 3.4% of tokens in Wikipedia
- DepthFive: 9.7% of tokens in Wikipedia

Once the corpora were created, they were dependency-parsed in the same way as the previous corpora and were used for extraction of IDPs with the POS-Dep system.

7.3.3 Wikipedia vs. specialised corpora

IDPs were extracted from Wikipedia, the G1M image caption corpus and the four alternative visual arts corpora created. Table 7.3 contains the 10 most frequently occurring IDPs extracted from each one of the six corpora for the image “Coney Island” (Figure 7.1) using the POS-Dep system. The following observations can be made from the table:

- The image caption corpus has produced IDPs with themes commonly found in images. For instance, there is a large number of IDPs describing scenes from a beach, which is not the case in Wikipedia. Although all IDPs look acceptable, this tendency towards triples from popular image themes could potentially be a disadvantage because less visually descriptive tags such as “popular” and “american”, even if important in the tag cloud, are given less prominence within the suggested IDPs.
- When using Visual Arts Wikipedia one can notice that IDPs containing words like “painting” and “figures” are prioritised. This becomes more obvious in the lower-depth versions of the corpus (e.g. DepthTwo or DepthThree). Although these corpora were created with tag disambiguation in mind (e.g. for tags like “stroke”), their main difference with respect to Wikipedia ended up being the focus on art vocabulary. This is not necessarily desirable, since the ‘focus’ of the tag cloud (expressed by its most popular tags) is typically not on the artistic characteristics of the image, but on particular objects and events depicted.

As can be seen, specialised corpora might, in principle, be useful for extracting IDPs, but they tend to favour a particular kind of vocabulary, which is not necessarily the one favoured by the tag cloud in question. Hence, I decided that using a general-purpose corpus like Wikipedia might be more suitable for my task. The size of Wikipedia, along with the wide range of topics it covers, allows for the extraction of a large number and variety of IDPs, which, given appropriate well-formedness and plausibility constraints, can constitute acceptable tag-relation-tag triples with respect to an image.

Table 7.3: The 10 most frequent Instantiated Dependency Patterns extracted using the POS-Dep system from six different corpora

Wikipedia		
<i>american</i>	\xleftarrow{ncmod}	<i>people</i> 2166
<i>oil</i>	\xleftarrow{ncmod}	<i>painting</i> 1921
<i>crowd</i>	\xrightarrow{ncmod}	<i>of</i> \xrightarrow{dobj} <i>people</i> 1395
<i>hot</i>	\xleftarrow{ncmod}	<i>summer</i> 597
<i>hot</i>	\xleftarrow{ncmod}	<i>balloon</i> 480
<i>sand</i>	\xleftarrow{ncmod}	<i>beach</i> 381
<i>american</i>	\xleftarrow{ncmod}	<i>pyramid</i> 301
<i>popular</i>	\xleftarrow{ncmod}	<i>beach</i> 245
<i>popular</i>	\xleftarrow{ncmod}	<i>figures</i> 240
<i>american</i>	\xleftarrow{ncmod}	<i>painting</i> 228
Image Captions (G1M)		
<i>sand</i>	\xleftarrow{ncmod}	<i>beach</i> 1235
<i>sand</i>	\xrightarrow{ncmod}	<i>on</i> \xrightarrow{dobj} <i>beach</i> 177
<i>beach</i>	\xleftarrow{ncmod}	<i>sand</i> 108
<i>sand</i>	\xrightarrow{ncmod}	<i>of</i> \xrightarrow{dobj} <i>beach</i> 56
<i>sand</i>	\xrightarrow{ncmod}	<i>at</i> \xrightarrow{dobj} <i>beach</i> 49
<i>hot</i>	\xleftarrow{ncmod}	<i>balloon</i> 41
<i>oil</i>	\xrightarrow{ncmod}	<i>on</i> \xrightarrow{dobj} <i>beach</i> 40
<i>beach</i>	\xrightarrow{ncmod}	<i>with</i> \xrightarrow{dobj} <i>sand</i> 24
<i>oil</i>	\xleftarrow{ncmod}	<i>painting</i> 23
Visual Arts Wikipedia (Depth-Five)		
<i>oil</i>	\xleftarrow{ncmod}	<i>painting</i> 1408
<i>american</i>	\xleftarrow{ncmod}	<i>painting</i> 182
<i>american</i>	\xleftarrow{ncmod}	<i>people</i> 168

<i>crowd</i> \xrightarrow{ncmod} <i>of</i> \xrightarrow{dobj} <i>people</i>	166
<i>oil</i> \xrightarrow{ncmod} <i>on</i> \xrightarrow{dobj} <i>painting</i>	161
<i>hot</i> \xleftarrow{ncmod} <i>balloon</i>	73
<i>hot</i> \xleftarrow{ncmod} <i>summer</i>	58
<i>figures</i> \xrightarrow{ncmod} <i>in</i> \xrightarrow{dobj} <i>painting</i>	35
<i>american</i> \xleftarrow{ncmod} <i>figures</i>	34
<i>popular</i> \xleftarrow{ncmod} <i>figures</i>	34

Visual Arts Wikipedia (Depth-Four)

<i>oil</i> \xleftarrow{ncmod} <i>painting</i>	713
<i>american</i> \xleftarrow{ncmod} <i>painting</i>	122
<i>oil</i> \xrightarrow{ncmod} <i>on</i> \xrightarrow{dobj} <i>painting</i>	87
<i>american</i> \xleftarrow{ncmod} <i>people</i>	77
<i>crowd</i> \xrightarrow{ncmod} <i>of</i> \xrightarrow{dobj} <i>people</i>	62
<i>hot</i> \xleftarrow{ncmod} <i>balloon</i>	34
<i>figures</i> \xrightarrow{ncmod} <i>in</i> \xrightarrow{dobj} <i>painting</i>	26
<i>painting</i> \xrightarrow{ncmod} <i>in</i> \xrightarrow{dobj} <i>oil</i>	16
<i>american</i> \xleftarrow{ncmod} <i>figures</i>	14
<i>popular</i> \xleftarrow{ncmod} <i>painting</i>	14

Visual Arts Wikipedia (DepthThree)

<i>oil</i> \xleftarrow{ncmod} <i>painting</i>	166
<i>american</i> \xleftarrow{ncmod} <i>painting</i>	56
<i>american</i> \xleftarrow{ncmod} <i>people</i>	20
<i>hot</i> \xleftarrow{ncmod} <i>balloon</i>	18
<i>crowd</i> \xrightarrow{ncmod} <i>of</i> \xrightarrow{dobj} <i>people</i>	11
<i>sand</i> \xleftarrow{ncmod} <i>painting</i>	11
<i>figures</i> \xrightarrow{ncmod} <i>in</i> \xrightarrow{dobj} <i>painting</i>	9
<i>figures</i> \xrightarrow{ncmod} <i>of</i> \xrightarrow{dobj} <i>people</i>	9
<i>oil</i> \xrightarrow{ncmod} <i>on</i> \xrightarrow{dobj} <i>painting</i>	8
<i>grotesque</i> \xleftarrow{ncmod} <i>figures</i>	7

Visual Arts Wikipedia (DepthTwo)

<i>oil</i> \xleftarrow{ncmod} <i>painting</i>	47
<i>american</i> \xleftarrow{ncmod} <i>painting</i>	14
<i>sand</i> \xleftarrow{ncmod} <i>painting</i>	7
<i>figures</i> \xrightarrow{ncmod} <i>in</i> \xrightarrow{dobj} <i>painting</i>	6

<i>oil</i>	\xleftarrow{ncsubj}	<i>predicts</i>	\xrightarrow{dobj}	<i>painting</i>	3
<i>outdoor</i>	\xleftarrow{ncmod}	<i>painting</i>			3
<i>painting</i>	\xleftarrow{conj}	<i>and</i>	\xrightarrow{conj}	<i>figures</i>	3
<i>painting</i>	\xleftarrow{ncsubj}	<i>portrays</i>	\xrightarrow{dobj}		3
<i>people</i>					
<i>american</i>	\xleftarrow{ncmod}	<i>people</i>			2
<i>figures</i>	\xrightarrow{ncmod}	<i>of</i>	\xrightarrow{dobj}	<i>people</i>	2

7.4 Summary

In this chapter, I explained my decisions regarding the design of a system that identifies semantically related tag pairs and postulates relations between them without utilising visual data or image-specific text. I confirmed the feasibility of such a task after demonstrating that it can be successfully completed by humans. In the rest of the chapter, I described two systems, Dep and POS-Dep, which are designed to extract relations that are both well-formed and plausible with respect to a given image. Relations were induced from Wikipedia, which was preferred over visual arts corpora and an image caption corpus. POS-Dep, which I chose as the most appropriate system, is evaluated in the next chapter.

Chapter 8

Evaluating postulated triples

Having described the design of a proof-of-concept system which aims to show the possibility of postulating well-formed and plausible tag-relation-tag triples, I will proceed to evaluating the quality of its output. In this chapter, I discuss evaluation decisions and explain evaluation experiments and their outcomes. In Section 8.1, I justify the evaluation measures I use, while in Section 8.2, I describe a baseline system, which will help quantify the contribution of the system proposed in this thesis. In Section 8.3, I discuss the appropriateness of two testbeds for evaluation, **i)** unseen data from the parallel corpus and **ii)** *a posteriori* human judgements, and explain how testing was attempted on the former. In Section 8.4, I describe human evaluation pilots and finally, in Section 8.5, I discuss details of the main human evaluation experiment along with the results obtained.

8.1 Measuring the quality of suggested triples

In order to determine the appropriate measures for evaluating the output of the proposed system, POS-Dep, it is important to understand what paradigm of systems it belongs to. Examining typical evaluation methods within such a paradigm can help inform evaluation choices for the purposes of this work.

Although POS-Dep performs a novel task, its structure can be compared to that of a typical search engine. As explained in Section 2.3, tagging is similar to indexing in the context of Information Retrieval (IR). Both practices aim to annotate a document in order to render it searchable in the future. A tagging system consists of an interface, typically web-based, that allows users to label resources with keywords, and is almost always part of a search engine which allows users to retrieve documents on the basis of tags assigned to them. For instance, a user of an image tagging platform can retrieve images whose tags are compatible with the user’s query. What differentiates a tag-based search engine from a typical search engine is the fact that the index terms of the former (i.e. tags), that help match a document to a query, have been crowd-sourced as opposed to automatically extracted. In both tag-based and typical search engines, we can distinguish four information retrieval stages:

1. establishing an **information need**, that is having a thought about what information to request. Information needs are known as ‘topics’ in the Text Retrieval Conference (TREC); see §2.3.1.

2. submitting an **information request** (query) to the search engine. A query is an interface through which an information need (topic) is articulated and communicated to the search engine. One topic can be expressed in more than one queries.
3. matching a query to **information assertions** (index terms) that have been made about a document. In a typical search engine, index terms are automatically extracted and weighted, with methods such as term frequency - inverse document frequency (Salton et al., 1975). In a tag-based search engine the index terms for a document are tags submitted manually by users and their weights are a function of the number of times that each tag has been used for the document (image in this case). As discussed in Section 2.3, index terms are connected with a document via an ‘aboutness’ relation; each index term provides some information regarding what the image is *about*. In fact, search queries themselves also hold an ‘aboutness’ relation with some images (i.e. the ones they are intended to retrieve).
4. retrieving a ranked list of **information sources** (documents).

The above stages are illustrated in Figure 8.1.

The POS-Dep system uses index terms, that is tags, to recreate thoughts that users could have had while annotating a given image. Such thoughts can also be seen as potential information needs satisfiable by a given image. Therefore, it can be claimed that the output of POS-Dep, that is Instantiated Dependency Patterns (IDPs), is a rudimentary representation of a topic. For instance, an IDP such as *house* \xrightarrow{ncmod} *in* \xrightarrow{dobj} *countryside* can be seen as the representation of an information need that the image in Figure 5.1a (page 64) satisfies.

Hence, POS-Dep performs advanced indexing whereby tags are enriched with proposed underlying relationships.

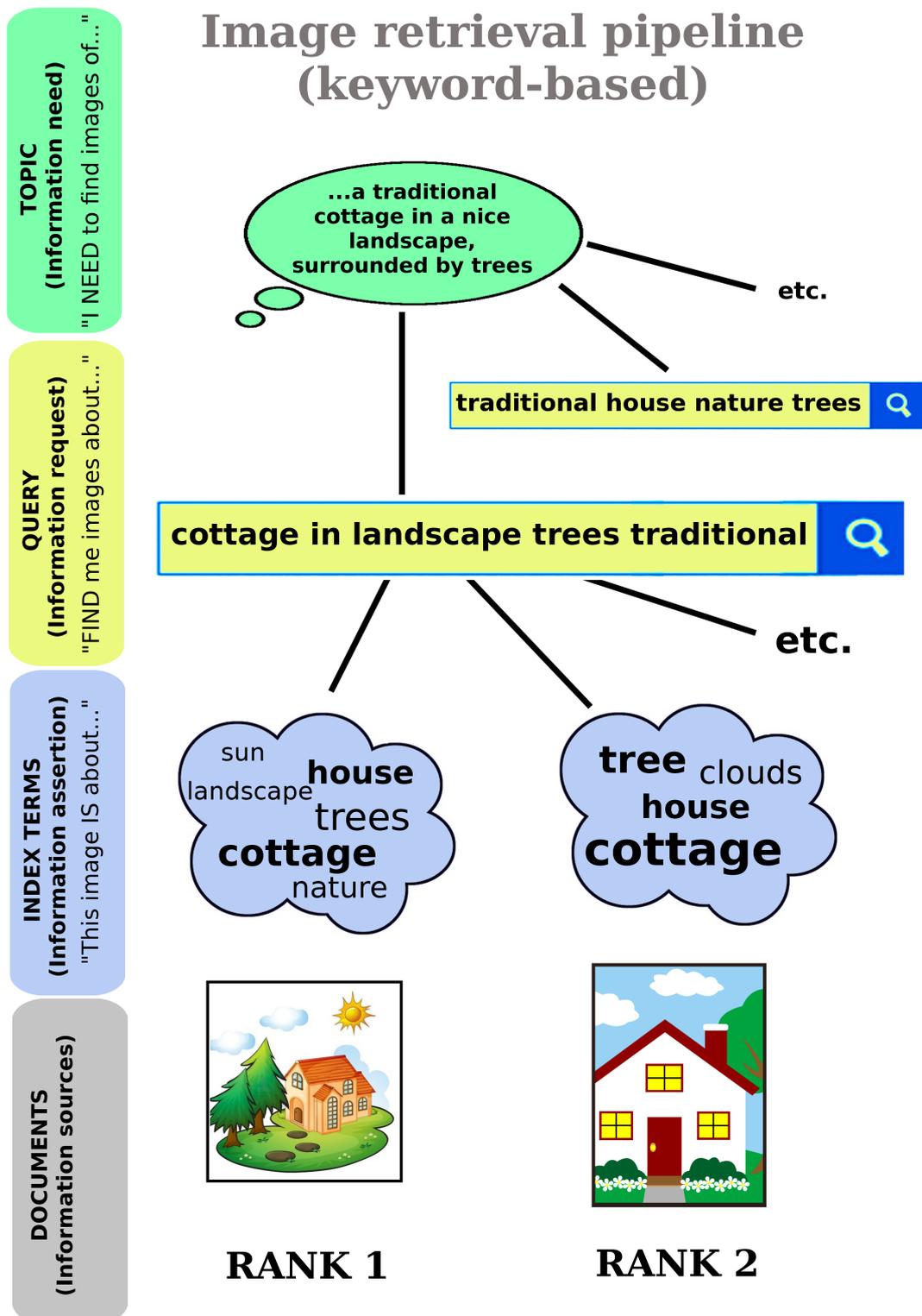


Figure 8.1: **Information layers in keyword-based image retrieval.** The four stages are referred to as 'Information needs', 'Information requests', 'Information assertions' and 'Information sources'.

Evaluation measures

If the ideal output of POS-Dep is a list of all and only acceptable relations holding between tags, then it would be natural to investigate whether typical IR evaluation measures, such as precision and recall, can be applied to its evaluation. In IR, precision is the percentage of correctly retrieved documents relative to all retrieved documents. Simply put, precision quantifies how correct the output is. Recall is the percentage of correctly retrieved documents relative to all correct documents, whether retrieved or not. Hence, recall quantifies how well the output covers the set of correct cases.

With respect to the output of POS-Dep, a measure similar to *precision* is obviously necessary: one needs to know how acceptable (§7.2) the generated topics (IDPs) are. *Recall*, on the other hand, is problematic: it is impossible to determine the complete set of topics satisfied by the image. Any attempt to estimate the set of all and only acceptable topics of an image cannot be assessed for reliability, since its size is simply unknown and potentially enormous.

Obtaining some acceptability score, similar to precision, might alone be very informative when attempting to quantify system performance. The importance of a precision-based scoring for IDPs can be further understood when POS-Dep is seen in the context of topic modelling (Steyvers and Griffiths, 2007) and topic labelling (Magatti et al., 2009; Lau et al., 2011; Hulpus et al., 2013; Aletras et al., 2014). Topic labelling tasks aim to produce labels – usually natural language strings – for topics of a document, themselves generated by a topic model. Like topic modelling, POS-Dep discovers topics and, like topic labelling, it represents them in easily understood format (IDPs); topic discovery and production of IDPs is performed by POS-Dep simultaneously. In both topic modelling and topic labelling systems, performance is measured with respect to the *interpretability* of the output (e.g. ‘Are topics or topic labels produced interpretable as a semantic unit?’); no attempt is made to estimate the extent to which the output covers the set of all possible cases, since this set cannot be known. Likewise, the IDPs produced by POS-Dep can be judged for *acceptability* (i.e. well-formedness and plausibility) with a measure similar to precision, which can provide enough reliable information on system performance.

Acceptability judgements

To determine the acceptability of the postulated IDPs, the evaluation module needs access to *acceptability judgements*, analogous to relevance judgements used in IR. In typical search engines, relevance judgements are binary decisions on whether a document from a fixed collection is relevant or non-relevant to an information need (Voorhees and Harman, 2005). Similarly, for POS-Dep, acceptability judgements should be decisions on whether an IDP is or is not acceptable as an information need satisfiable by the image. However, obtaining acceptability judgements for the output of POS-Dep is complicated, since there is no fixed collection of IDPs on which to make *a priori* human judgements. In other words, it is not possible to present human judges with a complete collection of potentially acceptable (or unacceptable) IDPs before the system has made its suggestions.

To compensate for the lack of a fixed IDP collection – and, hence, the lack of proper acceptability judgements – in advance of system output, I have performed evaluation on two different testbeds. The first testbed (unseen parallel subcorpus; §8.3.1) uses indirect *a priori* human judgements, inferred from the text submitted by the 68 participants in the unseen data of the parallel corpus experiment, which had been left aside for testing (see

beginning of chapter 6). These judgements are far from complete given their small number, therefore, this testbed is not considered a gold standard; testing against the unseen parallel subcorpus constitutes an initial sanity check before the beginning of (more costly) human experiments. The second testbed (§8.3.2) uses *a posteriori* human judgements, obtained from participants assessing system output.

8.2 Baseline

Whichever performance metric one uses to evaluate a system, results are easier to interpret if compared against some reference system, for instance, an existing state-of-the-art system that attempts to perform the same task. With respect to this research, there is, to my knowledge, no previous system designed to suggest acceptable image-specific tag-relation-tag triples. Hence, the benchmark against which performance will be evaluated is a baseline, that is a simple system attempting to solve the same problem in a less sophisticated way. All examples provided in this section are for the image “Detroit 1943” (page 105).

The baseline used in this work is called the ‘Ngram’ system. For each image, this system extracts strings of words from Wikipedia as bigrams, trigrams and fourgrams whose leftmost and rightmost words are tags in the image’s tag cloud (e.g. “peaceful landscape”, “house in countryside”, “house surrounded by trees”). Determiners were ignored during the construction of the ngrams in an attempt to facilitate comparison between the baseline and the proposed system, which extracts IDPs through grammatical dependencies that ‘jump over’ determiners. To give an example, for a pair of tags such as “people” and “city”, the Ngram system extracts the trigram “people in city” and not the fourgram “people in the city”. By omitting determiners, this system allows for 4-word phrases such as “people walking in city”, which are comparable to the ones extracted by POS-Dep. The reason why determiners do not exist in the IDPs extracted is that words with this part of speech tend to be connected with only one element (a noun) in the dependency graph, failing to occur in the middle of an ‘unbranched catena’ (see §6.3.1), which forms an IDP.

Ngrams lack the syntactico-semantic information that dependencies provide, since they comprise a simple string of words. However, comparison of ngrams with the IDPs from the POS-Dep system is legitimised by the fact that ngrams often contain meaningful word combinations which may be parseable into dependency structures. For instance, the trigram “people in city” is well-formed in English and captures a plausible relation between the tags “people” and “city”; if it were to be parsed, it would form an IDP such as $people \xrightarrow{ncmod} in \xrightarrow{dobj} city$.

Through informal observation, it can be seen that ngrams with high counts, especially bigrams (e.g. “old buildings”) and trigrams (e.g. “black and white”) are often not only well-formed but also plausible with respect to a given image. Plausibility can be explained by the fact that frequently occurring ngrams can capture frequent semantic relations (“people” and “cars” are often connected with a relation denoting driving), which can render the ngram (e.g. “people driving cars”) plausible for a wide range of images. The inherent well-formedness and plausibility of frequently occurring ngrams makes the Ngram system a strong baseline. In addition, Ngram is expected to be strong at discovering pairs of related tags, which is part of the task that POS-Dep is performing (see §7.1.1). Higher ngrams, especially fourgrams, can be noisy, with irrelevant words interfering (e.g. “street

in limerick city” extracted for tags “street” and “city”; “white shirt with black” for tags “white” and “black”).

To enable comparison between Ngram and POS-Dep, I converted the IDPs of the latter into strings of concatenated nodes, that is words. For example, $people \xleftarrow{ncsubj} live \xrightarrow{ncmod} in \xrightarrow{dobj} city$ becomes “people live in city”. If more than one IDP results in the same string, only the first copy of the string is kept, that is the one created from the most frequent underlying IDP. Although converting IDPs to strings involves some information loss (since converting back from strings to IDPs is not deterministic), this loss is minimal: in the rare cases where a string has more than one underlying IDP, the preferred IDP is easy to predict. For example, the string “oil painting” is 320 times more likely to derive from $oil \xleftarrow{ncmod} painting$ than from $oil \xleftarrow{ncsubj} painting$ in Wikipedia; the latter is the result of mis-parsings.¹ From now on, these strings will be referred to as *phrases*; the term is used to describe a stretch of words, without implying the existence of a grammatical phrase, such as a noun phrase.

A random baseline could also be used as a benchmark. For instance, the output of POS-Dep could be compared to that of a system that matches tags at random (e.g. by sampling from the tag cloud) and connects them with randomly selected words from a corpus. However, whichever the evaluation measure, such a baseline would essentially have zero scores, hence it was not considered a useful benchmark for the proposed system.

8.3 Testbeds

8.3.1 Unseen parallel subcorpus

Evaluating a system against a corpus of *a priori* judgements can be beneficial because, unlike time-consuming human evaluation, it can allow for multiple measurements that might help guide system design or reveal interesting aspects of system performance. Such a corpus, when large and created after multiple annotators have reached consensus, is regarded a gold standard: the more a system’s output resembles it, the better this system is said to perform. However, as mentioned in Section 8.1, a gold standard could not be created for evaluating POS-Dep, since a corpus like this would require acceptability judgements on all and only correct IDPs that could possibly apply to each image. Such a set is unrealistic to obtain. As an intermediate solution, experiments could be conducted on the unseen data (tags and text from 68 participants) that had been left aside after the construction of the parallel corpus.

To obtain IDPs for the benchmark, against which ‘Ngram’ and ‘POS-Dep’ will be evaluated, I used the tags and textual descriptions of the unseen subcorpus for each one of the five images involved (Figure 5.1; page 64). Benchmark IDPs were extracted if they could unify with the Abstract Dependency Patterns (§6.3.2) learnt from the larger (seen) parallel subcorpus (150 participants). For the sake of this experiment, all benchmark IDPs obtained for each picture were considered *acceptable*, while any of the possible IDPs that were not found in the participants’ text were considered *non-acceptable*. This is, obviously, a strong assumption: although we can be confident that the IDPs obtained from the unseen corpus are *only* (or to a great extent) acceptable, by analogy to those of the examined 150-participant subcorpus, it is very likely that these are not *all* possible

¹The first IDP occurs 1,921 times while the second occurs only 6 times.

acceptable IDPs. Each one of the five images in the unseen subcorpus has an average of 120 IDPs², which seem too few compared to the approximately 13,000 IDPs per image obtained by the (even more restrictive) POS-Dep system from Wikipedia. Therefore, results of this experiment should only be taken as some initial indication of system performance before evaluation against the more reliable *a posteriori* human judgements.

This experiment compares the performance of POS-Dep against that of Ngram using phrases (e.g. “house in countryside”) as output and precision as evaluation measure. The benchmark IDPs were also converted to phrases.

Precision for each system was equal to $TP/(TP + FP)$, where TP is the number of true positives, FP is the number of false positives, thus, $TP + FP$ is the total number of phrases produced by the system. Correct phrases are those that are found in the unseen parallel subcorpus; wrong phrases are those not found. We can assume that many more acceptable IDPs exist than those found in this small-scale corpus. Hence, it is expected that precision of POS-Dep output will be underestimated.

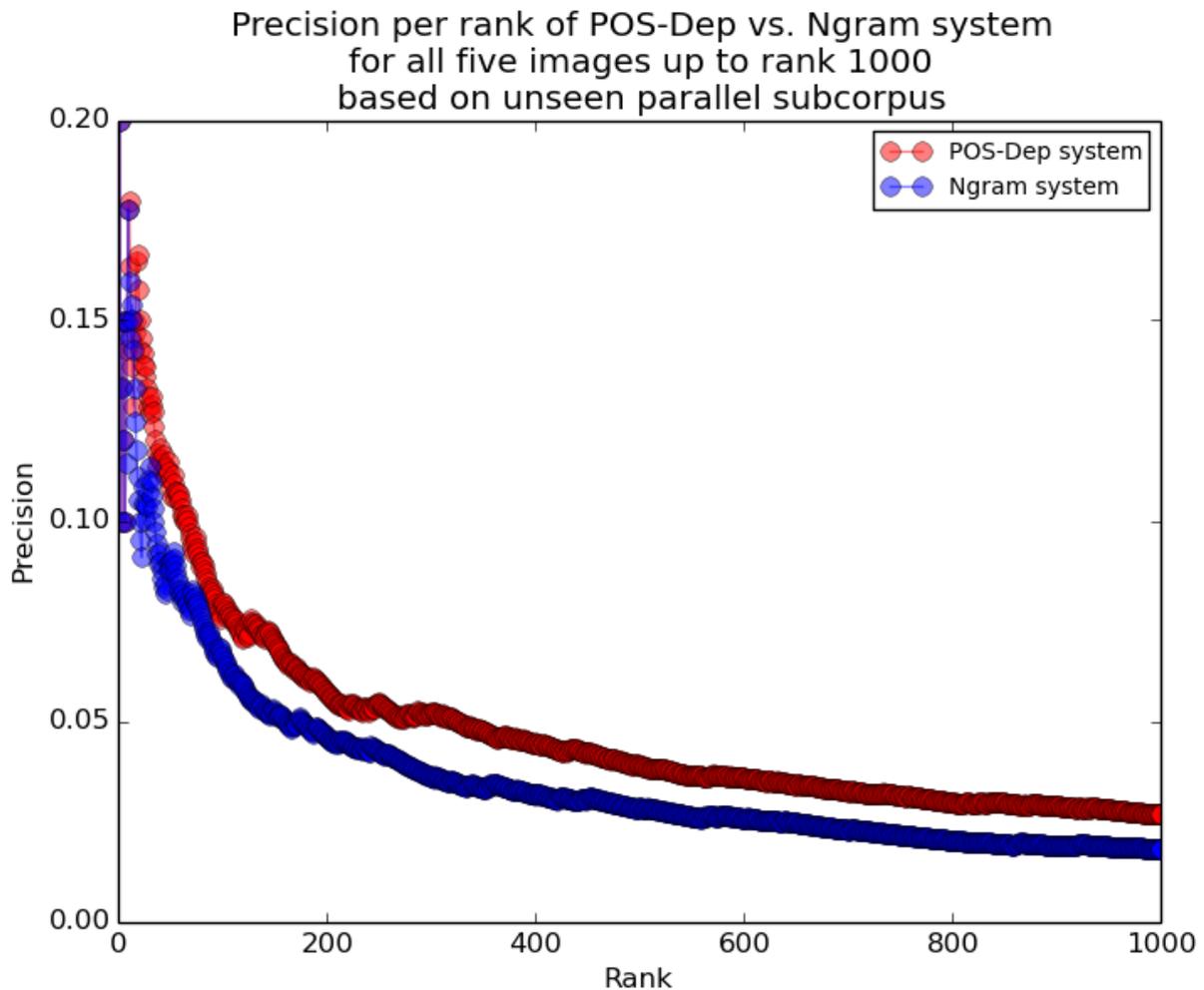


Figure 8.2: Comparison of POS-Dep and Ngram system performance based on unseen parallel subcorpus (68 participants)

²143 for “House in Provence”, 128 for “Torso”, 96 for “Angel of Resurrection”, 99 for “Moulin Rouge: La Goulue” and 134 for “The Two Sisters”

Figure 8.2 is a plot of the precision scores for POS-Dep and Ngram for the top 1,000 ranks, when rank is determined by frequency in the corpus. This initial result shows that the POS-Dep system has a larger area under the curve (AUC) than the baseline. Although the two systems perform similarly in the first few ranks, probably because of high-frequency ngrams (especially bigrams), POS-Dep clearly outperforms Ngram. The results of this initial experiment legitimise the next step of the evaluation process, which involves more time-consuming experiments with humans. In terms of absolute numbers, precision of both systems is low, however, this is a side-effect of the corpus being small since absent phrases are marked as non-acceptable.

8.3.2 A posteriori human judgements

Since *a priori* judgements available for the evaluation of POS-Dep are far from complete, more reliable results can be obtained with *a posteriori* human evaluation. The system’s output, a list of topics (information needs) represented as IDPs, can be assessed for acceptability not through comparison to a reference corpus but through humans evaluating the output itself. However, if the postulated IDP themselves (e.g. *vase* \xrightarrow{ncmod} *with* \xrightarrow{dobj} *flowers*) were presented to humans for evaluation, some level of expertise on the part of the assessors would be necessary. For instance, such a task would require an understanding of the semantics made explicit in a dependency structure (i.e. functor-argument relations hinted by heads and dependents), which would make it difficult to recruit evaluators. Presenting IDPs to humans would also inhibit comparison of the system’s performance against the baseline, whose output is simple strings (e.g. “buildings are old”).

To overcome the complications arising from using IDPs for *a posteriori* human judgements, I decided to present humans with phrases extracted from IDPs, as done in the initial evaluation experiment (§8.3.1). The assumption underlying this decision is that phrases can preserve enough information for a judge to decide on the acceptability of the underlying IDP. If the phrase is *idiomatic* (i.e. natural-sounding English), then it is very likely that the IDP that it derives from is well-formed; if it is not idiomatic (e.g. ungrammatical or a bad collocation), then its underlying structure is probably flawed. Similarly, if the phrase makes a *plausible* statement about the image, then it is very likely that the IDP is also plausible.

A legitimate question to ask at this stage is: Why are IDPs extracted if they are not used in any evaluation experiment? What justifies the extraction of deeper relations? The answer is that POS-Dep phrases given to evaluators could not have been produced without any linguistic analysis of the sentences they derive from. For instance, POS-Dep phrases tend to omit noisy words from relations (e.g. “house in countryside” is produced but “house in french countryside” is not) since the originally extracted IDPs ‘jump over’ words to form unbranched catenae. Hence, despite the small information loss that conversion from IDPs to phrases involves, evaluation of the latter can still be used to draw conclusions on the former.

Before asking humans to assess phrases for acceptability, it is important to examine ways in which judges should be instructed to make this decision. For instance, asking them whether they consider a given phrase “well-formed” (or “idiomatic”, “natural sounding” etc.) and “plausible with respect to the image” (or “truthful”, “relevant” etc.) might be either a clear and simple task presentation, or too vague a request, resulting in a high

degree of subjectivity. Whether a simple task like this is possible for humans to perform reliably is investigated in the first pilot experiment (§8.4.1).

A more detailed evaluation task that humans could be instructed to complete involves assessing the quality of the output with respect to a real-world scenario. Since system performance is evaluated through phrases, a real-world evaluation task can be chosen among the contexts in which such phrases act as utterances. Below I discuss three possible linguistic functions that this type of phrases could fulfil if uttered: **i)** phrases as search queries, **ii)** phrases as descriptions and **iii)** phrases as elliptical statements.

Phrases as search queries Since tagging aims to facilitate future retrieval of a document, it would be natural to consider treating the phrases as potential search queries, and evaluate them in a search-based scenario (e.g. asking judges whether they would use such ‘queries’ to retrieve a particular image). However, such phrases are not structurally similar to search queries. A string such as “cottage in nature trees traditional” constitutes a reasonable search query but is not parseable as a single catena, syntactic phrase or as any other part of a hypothetical underlying sentence. Often, only a subset (or subsets) of the query consists of syntactically linked words. It is also common to see queries with words that have no obvious syntactic relations with each other (e.g. “spain flag”) but with a clear semantic relation. Roy et al. (2014) suggest that web search queries represent a ‘protolanguage’, with a mix of elements that may or may not resemble the structure of natural language. As the authors observe, this simple language is gradually shifting towards more complexity, that is it starts resembling natural language more through the years, possibly because of the increasing ability of search engines to match syntactically connected units to relevant documents. However, at present, queries are still written in this semi-structured language, which is *less complex* than the phrases that the following evaluation experiments aim to assess.

Phrases as textual descriptions Alternatively, phrases could be treated as textual descriptions of the image. Describing an image with words can take different forms in different contexts. For example, image description can be equivalent to captioning, which often includes comments on an image, that is complementary information, not directly observable in the picture (see §7.3.1). Textual description of an image can also be used to ‘translate’ a visual stimulus into language for situations when the stimulus is not available (e.g. describing the image to someone who cannot see it; see §5.4.2). Furthermore, image description can be equivalent to producing an in-depth analysis of an image or its medium (e.g. a historical account and formal characteristics of a painting). However, what all such situations have in common is that descriptions tend to comprise full sentences. Hence, textual description as a linguistic function is *too complex* to be performed by phrases used in this research.

Phrases as elliptical statements Many of the phrases derived from IDPs can be seen as elliptical grammatical phrases (e.g. “beach under sun”), which ‘jump over’ words found in fully fledged sentences. Such constructs can only constitute plausible utterances when they perform special functions. Stainton (2006) suggests that elliptical constructions are the norm within ‘special registers’, such as recipes, telegrams, newspaper headlines, diaries, note taking and text messaging, or within ‘protolanguages’, such as child languages and pidgens. Some of these functions (e.g. recipes) would be difficult to apply to IDP-

derived phrases, which are given with respect to an image. However, real-life scenarios could be devised for many other functions of phrases as elliptical statements. Presenting phrases to human assessors in a note-taking scenario was explored in the last two evaluation experiments (§8.4.3 and §8.5).

8.4 Pilot experiments

Given that *a posteriori* human judgements were necessary for assessing the performance of POS-Dep, a series of evaluation experiments were conducted to this end. This section describes three pilot experiments whose aim was to provide feedback on aspects of evaluating the suggested tag-relation-tag triples using human judges, thus preparing the ground for the main experiment (§8.5).

The pilot experiments took place over the course of three months during system construction. Since the purpose of the pilots was not to assess the quality of the output but to test parameters of the evaluation process, it was considered unnecessary to delay piloting until the system was finalised. The three pilots were conducted one after the other, approximately one month apart. In each pilot, phrases were presented to humans from a different version of the system under development.

Participants for the pilot studies, as well as for the main experiment, were University of Cambridge postgraduate students. Recruitment was conducted through personal email containing a web link that the participant could follow to start the experiment. For each experiment (pilot or main), an email was sent to a different group of people. This ensured that each participant had not assessed system output in the past. Approximately one third of the people contacted for the pilot studies had participated in pilot experiments for the construction of the parallel corpus (see Chapter 5). However, for the main evaluation experiment, all people contacted were completely unfamiliar with previous studies. The first and the third pilot (Sections 8.4.1 and 8.4.3) as well as the main experiment (Section 8.5) were conducted online, while the second pilot (Section 8.4.2) was performed face to face. With the exception of the second pilot, the identity of each participant who completed an experiment was not recorded. No demographic data was collected, since analysing responses with respect to participant variables was outside the scope of the evaluation experiments.

8.4.1 Pilot experiment 1

So far, acceptability judgements have been treated as *binary*. For instance, in the initial evaluation experiment (§8.3.1), a phrase from POS-Dep was marked as acceptable if it appeared in the unseen corpus and as non-acceptable otherwise. However, binary judgements might be difficult for a human to make if phrases are acceptable to *varying* degrees. This pilot experiment was primarily designed to measure whether participants have difficulty deciding on the acceptability of phrases when faced with a binary choice. A secondary objective was to receive feedback on the clarity of the instructions.

Through a purpose-built online interface, participants were presented with phrases, derived from IDPs which had been extracted from Wikipedia. The IDPs had been extracted using only ‘path constraints’ (see §6.3.1, p. 85), and not ‘pattern constraints’ (§6.3.2, p. 89). The reason was that such a configuration produces mixed-quality output,

which can better illustrate the participants' difficulty (or ease) of deciding on a yes-or-no basis.

In this experiment, I aimed to collect judgements for approximately 150 phrases. Acknowledging that such a number of phrases might impose a heavy cognitive load on any single participant, I divided phrases in two sets and presented them to two separate groups of people, so each individual would be required to evaluate no more than 75 phrases. Having two groups of participants enabled the evaluation of a large number of phrases and, at the same time, left enough subjects in each group for the measurement of inter-rater reliability, which would quantify the participants' consensus and, by extension, their ability to make binary decisions.³

Each group was shown the same six images but different phrases. Three of the visual stimuli were: **i)** "The Grizzly Giant Sequoia" (Figure 7.8b, page 105), **ii)** "Detroit, 1943" (Figure 7.8a, page 105) and **iii)** "Coney Island" (Figure 7.1, page 98). The other three can be seen in Figure 8.3.



(a) "The Cotton Pickers" (1876) by Winslow Homer



(b) "Loss of the Schooner 'John S. Spence' of Norfolk, Virginia" (1833) by Thomas Birch



(c) "Bacchus and Ariadne" (1754) by François Boucher

Figure 8.3: Three of the five images used in the first pilot evaluation experiment

³For the evaluation experiments, the term 'inter-rater reliability' is preferred to the more commonly used term 'inter-annotator agreement' since the latter refers to agreement between annotators (i.e. raters labelling data, which may be used for training a system).

Each individual was presented with 72 phrases (12 phrases in each of the six images). In the two participant groups, I attempted to represent **i**) as many underlying dependency patterns as possible and **ii**) all phrase lengths (two, three and four words) equally, by sampling in a way that does not disadvantage less frequently occurring constructs.⁴ The order of the images was differentiated (see §5.4.3). Next to each image, the 12 phrases shown were randomised.

Metadata, such as the title of the artwork shown in the image and the artist, were visible. The tags submitted for the images in the Steve corpus were not made available to the participants. Screenshots of the initial instructions as well as a sample page shown to participants can be found in Figures B.1 (page 183) and B.2 (page 184).

Outcome The experiment was completed by 13 participants (seven in Group One and six in Group Two). To quantify the subjects’ ability to make binary relevance judgements, I measured inter-rater reliability (henceforth IRR) between the participants of each group. I used Fleiss’ kappa (Fleiss, 1971), which is suitable when variability of responses is measured between *more than two* participants and response categories are *binary*. The initial concern that users may have difficulty making binary judgements on the phrases was confirmed: kappa value was 0.014 for Group One and 0.002 for Group Two, which indicated no above-chance agreement. This result led to the realisation that binary relevance judgements were hard for humans to make. The complete responses can be found in Section B.1.3.

Before submitting their responses, users had the option of making comments on the experiment. Feedback received can be summarised as follows. The exact comments can be found in Section B.1.2.

- It was difficult for some participants to decide if a phrase is ‘good’ or ‘bad’; middle options should be considered.
- Details of particular images were not discernible on the screen.
- Some phrases were judged as non-acceptable based on the participant’s historical knowledge or personal opinion.
- There was some uncertainty over the status of phrases that are plausible but not well-formed. Participants commented that they considered these phrases ‘bad’, as per the original instructions.

In response to the above IRR and comments, the following action was taken:

- Graded, as opposed, to binary categories were considered for the experiments to follow.
- Instructions became clearer.
- Participants were discouraged from projecting their own opinion and, instead, were asked about the acceptability of a phrase based on a real-life scenario with a fictional character involved. This was implemented in Pilot 3 and the main experiment.

⁴For instance, sampling naïvely from the pool of phrases would produce mostly two-word phrases from the pattern $* \xleftarrow{ncmod} *$ (e.g. ‘green forest’), which accounts for more than 1/4 of phrases in each picture, or three-word phrases from the pattern $* \xleftarrow{conj} * \xrightarrow{conj} *$ (more than 1/5 of phrases).

- Hard-to-interpret images were avoided.

8.4.2 Pilot experiment 2

The second pilot evaluation experiment was conducted with graded, as opposed to binary, categories. It was undertaken by one participant, whose physical presence was required at the site of the experiment. The purpose of this pilot was to observe the subject's behaviour in real time and receive instant feedback on the difficulty, or ease, of making decisions and on the clarity of the instructions.

The participant was presented with a paper copy of the image “Coney Island” (Figure 7.1, page 98) and was asked to evaluate a list of 163 phrases, extracted from IDPs that occur at least four times in Wikipedia.⁵ The IDPs were extracted using an early version of the ‘Dep’ system (§7.2).⁶ No tags or metadata were provided, in an attempt to avoid further distractions. The task was to decide the extent to which each one of the given phrases could be used to describe the picture. The instructions were as follows: “Below is a list of expressions proposed for the image. On a scale from 0 to 10, to what extent do you think that someone could use such an expression to describe the image? Please indicate your answer next to the proposed expression.”

The full list of phrases with the participant's responses can be seen in the Appendix (§B.2, page 188). After completing the experiment, the participant was shown the tags and metadata. At the end of the experiment the participant commented that:

- providing a degree of confidence in a particular phrase was easier than having to make a binary decision as to its quality.
- it would be easier if two different types of questions had been asked: **i)** “To what extent would you use these expressions?”, **ii)** “To what extent do you think someone would use these expressions?” Answers would differ under each question. For example, the participant would not use the phrase “figures of people” to describe the image, but believes that an art historian would.
- if the participant had been shown the metadata before the evaluation, responses would have been different. For example, if the official title of the painting (“Coney Island”) had been shown, the confidence in all phrases starting with “american” would be higher.
- more detailed instructions are necessary, since the question of whether a phrase is a good descriptor of an image is rather vague.

In response to the above comments, the third pilot experiment **i)** provided a profile of the person supposed to have produced the phrases accompanying an image, **ii)** was designed after a more informed decision about the availability of tags and metadata to the participant, **iii)** provided more detailed instructions, that linked the evaluation task to a real-world scenario.

⁵For this experiment, the October 2010 Wikipedia dump was used; page accessible through WayBack Machine on <http://web.archive.org/web/20110520111534/http://dumps.wikimedia.org/enwiki/20101002/>

⁶As explained in the beginning of Section 8.4, pilot experiments were conducted while the inter-tag relation system was under development, hence the use of ‘Dep’ and not ‘POS-Dep’ in this study. Since the pilot experiments was to test aspects other than the quality of the output (e.g. instructions), piloting was not delayed until the system was finalised.

8.4.3 Pilot experiment 3

The third pilot experiment was conducted for the purpose of measuring IRR when relevance judgements are graded, as well as deciding on the right instructions. A high IRR would provide reassurance that the evaluation task is natural and straightforward, so there are no misunderstandings on the part of the participants, and that any results obtained from the assessors' responses are to be trusted as containing minimal statistical error, that is variance.

The aim of this pilot experiment was not to measure system performance, so the phrases assessed were not the best produced. On the contrary, there was an effort to select phrases that represent a wide range of qualities, lengths and underlying dependency patterns, on the premise that such a balanced set of phrases can reveal more tendencies regarding IRR. The varying qualities were due to the fact that the phrases evaluated were produced without 'pattern constraints', as per Pilot 1. The IDPs from which the phrases derived were sampled among those IDPs that occur at least four times in the October 2013 Wikipedia dump.⁷ Sampling was performed in a way that represents different phrase lengths and different underlying dependency patterns as 'democratically' as possible, but not proportionally. In other words, less frequent phrase lengths or patterns were sampled from as often as more frequent ones, as long as there were enough distinct IDPs to choose from.

Following the feedback received in the first two pilot experiments, a phrase was no longer assessed as either a 'good' or a 'bad' descriptor for an image, but was evaluated on its degree of acceptability. Acceptability was decided based on two explicitly stated criteria: **i)** well-formedness; the phrases had to be rated according to how idiomatic they were for the linguistic function they were supposed to fulfil, **ii)** plausibility; the phrases had to be rated according to how truthful they were with respect to the image. Since only one rating could be provided per phrase, assessors were instructed, through examples, to provide a low rating for phrases that fail to fulfil even one of the two requirements adequately, and to provide to high rating for phrases that fulfil both requirements to a good extent.

The task that participants were asked to perform was to decide the degree to which each one of the phrases could constitute a good 'note' (aide-memoire, in particular) for their respective images. The task was linked to a real-world scenario with the instructions:

Imagine the following scenario: Mary attended a talk at an art conference and saw some art pictures in posters. She didn't know the titles of the pictures but she wanted to find them later, so she decided to write down some short phrases as memory aids. You will be asked to guess to what extent you believe Mary could write each one of the phrases to remember the pictures: **1.** *Highly Unlikely*, **2.** *Probably Not*, **3.** *Cannot Decide*, **4.** *Probably*, and **5.** *Highly Likely*. Along with the phrases and the picture, you will be given a list of words that came to Mary's mind when she saw the image. These words might be useful when you guess what Mary ended up doing.

⁷<http://web.archive.org/web/20131212071039/http://dumps.wikimedia.org/enwiki/20131001>

The words that “came to Mary’s mind” were the tags attached to the image in the Steve corpus, excluding the ones that had not been used as end words for the phrases shown to the evaluators. The decision to include tags was made in an effort to discourage participants from projecting their own opinion on the image and encourage them to guess what the creators of tags would be saying if they were to use notes. One concern was related to the fact that the tags of a given image on the Steve website had been produced by more than one user, whereas in this experiment they were treated as a single person’s ‘thoughts’. However, this is not problematic if we consider the totality of tags on an image as reflecting the ‘public opinion’, which in this experiment was disguised as ‘Mary’.

Metadata about the image (e.g. information about the artist, museum and so on, as they appear on the Steve website) was removed on the grounds that: **i)** The participants might get confused by the increased amount of information that they need to use in order to make a decision, **ii)** The participants need to focus on what Mary thought, based on her tags and avoid the temptation of using the metadata to decide on their own. After all, the tags on the Steve tagging platform had been produced with metadata being present, so by showing tags to the participants, the instructions have already included ‘thoughts’ generated after users (here ‘Mary’) had considered the metadata. In the phrases suggested, the end words (i.e. tags) were underlined.

Stimuli The images shown to participants were four out of the six that had been used in the binary judgements experiment (§8.4.1). The two images that were omitted were “Bacchus and Ariadne” (Figure 8.3c, p. 129) and “Detroit, 1943” (Figure 7.8a, p. 105). The former was identified by participants as difficult to discern on a computer screen and the latter was dis-preferred in favour of pictures with less ambiguous themes.

Process The responses were collected through an online script. The welcome page and a sample evaluation page can be found in the Appendix (§B.3.1, page 190). The order of the images was differentiated for each participant (see §5.4), and the order of the phrases for each image was randomised.

Outcome The experiment was completed by three participants: two University of Cambridge students not previously exposed to the experiment, and myself. The inclusion of myself as one of the participants was motivated by the need to establish whether my intuitions regarding system output (which have guided decisions such as what system to include in the main evaluation experiment), agree with those of other assessors. Complete responses can be seen in Section B.3.2 (page 191).

Inter-rater reliability Each phrase was assessed by three judges, thus, its distribution of responses (i.e. ratings) is a vector of length 3, which I will be calling a *response vector*. For example, if the phrase “sisters in dresses” has been given a score of 3 by the first participant, a score of 4 by the second participant and a score of 2 by the third participant, then the response vector for this phrase is [3, 4, 2]. Since rating is performed on a 5-point Likert scale (Likert, 1932), which means that the categories assigned by assessors are ordinal, the IRR metric used has to allow for “degrees of disagreement”, as described in (Artstein and Poesio, 2008). A standard method in such cases is using Krippendorff’s α (Krippendorff, 1980), which treats each category – in this case each response vector – as a separate ‘level’ (e.g. ‘singers’ vs. ‘teachers’ vs. ‘athletes’) of a

single-factor (e.g. ‘occupation’) Analysis of Variance (ANOVA). ANOVA decides whether three or more samples (levels) are drawn from the same population, that is whether they are similar enough. For instance, the amount of sugar consumed per year might be found to be different among singers, teachers and athletes. By analogy, Krippendorff’s α aims to decide whether responses from three or more participants are similar enough. This measure is based on the ratio $\frac{s_{Within}^2}{s_{Total}^2}$, where s_{Within}^2 is the variance *within* the levels (here response vectors, whose variance is higher when participants disagree with each other) and s_{Total}^2 is the *total* variation of all levels considered together as one sample. When this ratio is equal to 0 (i.e. when $s_{Within}^2 = 0$), there is no variance within the levels (‘response vectors’), so there is perfect agreement between the annotators. When the ratio is equal to 1, then any agreement is due to chance. Finally, when the ratio is larger than 1, there is systematic disagreement. Krippendorff’s α is equal to 1 minus this ratio, so when $\alpha = 0$, there is chance agreement, when $\alpha > 0$, there is above-chance agreement, with $\alpha = 1$ indicating perfect agreement. If α is negative, then participants agree less than would be expected by chance.

Below is an example of three response vectors (‘items’), i_1, i_2, i_3 and i_4 a, submitted by three participants (‘coders’), c_1, c_2 and c_3 .

	c_1	c_2	c_3
i_1 : ships caught in storm	4	5	5
i_2 : sailors lost at sea	5	2	5
i_3 : rescue ships	4	2	3
i_4 : wind and waves	3	4	3

The mean of each response vector (item) is given by:

$$\bar{x}_i = \frac{\sum_{c=1}^C x_c}{C} \quad (8.1)$$

where C is the number of coders (i.e. assessors). The sum of squares within response vectors is given by:

$$SS_{within} = \sum_{i=1}^I \sum_{c=1}^C (x_{ic} - \bar{x}_i)^2 \quad (8.2)$$

where I is the number of items (i.e. levels, or response vectors) and x_{ic} is the value of item i provided by coder c . Degrees of freedom for calculating variance *within* the levels is found from:

$$df_{within} = I(C - 1) \quad (8.3)$$

Finally, error variance (s_{within}^2) is given by the following equation:

$$s_{within}^2 = \frac{SS_{within}}{df_{within}} \quad (8.4)$$

The grand mean of all responses from all participants is equal to:

$$\bar{x} = \frac{\sum_{i=1}^I \sum_{c=1}^C x_{ic}}{IC} \quad (8.5)$$

The total sum of squares is equal to:

$$SS_{total} = \sum_{i=1}^I \sum_{c=1}^C (x_{ic} - \bar{x})^2 \quad (8.6)$$

The degrees of freedom associated with calculating total variance are given by:

$$df_{total} = IC - 1 \quad (8.7)$$

Total variance can be calculated as follows:

$$s_{total}^2 = \frac{SS_{total}}{df_{total}} \quad (8.8)$$

Finally, Krippendorff’s α is equal to:

$$\alpha = 1 - \frac{s_{within}^2}{s_{total}^2} \quad (8.9)$$

Results IRR was greatly improved compared to the binary judgements experiment (§8.4.1), probably due to the introduction of the 5-point scale. Overall above-chance agreement was 0.47. When IRR was measured separately for each image, the results were: “Loss of Schooner” (0.63), “Grizzly Giant Sequoia” (0.56), “Cotton Pickers” (0.46) and “Coney Island” (0.16). Image “Coney Island” demonstrates that agreement may be problematic for particular images; a potential explanation might be the ambiguity of the themes described in the image. When IRR was reported separately for phrases of a particular length, the results were: length 2 (0.46), length 3 (0.50), length 4 (0.39). Given the subjective and possibly unnatural nature of the evaluation task, this level of IRR was considered adequate for drawing some tentative conclusions about the quality of the suggested inter-tag relations.

Feedback One of the participants commented that underlining end words in the phrases (e.g. ‘ships lost at sea’) distracted them from assessing the phrase’s idiomaticity. In response to this, tag underlining was removed in the main experiment.

8.5 Main experiment

The main evaluation experiment was conducted for the purpose of measuring and benchmarking the performance of POS-Dep using *a posteriori* acceptability judgements.



(a) “House in Provence” (1885) by Paul Cézanne



(b) “Loss of Schooner 'John S. Spence' of Norfolk, Virginia” (1833) by Thomas Birch



(c) “Sunlight” (1909) by Frank Weston Benson



(d) “Proserpine” (1874) by Dante Gabriel Rossetti

Figure 8.4: Visual Stimuli and their ids

8.5.1 Process

Image selection

Four images were presented to evaluators, all of which had been manually selected among the top 50 in the Steve corpus with respect to their total number of distinct tags. Preference was given to unambiguous and conceptually easy visual stimuli, since there was some indication from the previous experiment that IRR may be higher in images that convey a clear message. The images can be seen in Figure 8.4. Image “House in Provence” (Figure 8.4a) was among the images that had been used for the parallel corpus experiment (see §5.4). For this particular image, human-generated phrases (i.e. benchmark phrases in §8.3.1) were also included to help determine the extent to which these phrases can be distinguished by assessors from machine-generated phrases.

Phrase generation

Phrases were produced from POS-Dep and the Ngram baseline using Wikipedia as a corpus and eliminating IDPs that occur less than 6 times. For each image, both systems produced three rankings of phrases: one ranking for 2-word phrases, one for 3-word phrases and one for 4-word phrases. The reason was that, given the cost of human evaluation, only the top phrases per image would be assessed; a single ranking would disadvantage 3- and 4-word phrases, which tend to occur lower in the ranks, when ranking is determined – partly or entirely – by corpus frequency. In this experiment, ranking was based on a combination of criteria, including corpus frequency and cosine similarity of the phrase’s distributional vector to the tag cloud (see Section 7.2.2). For each one of the three rankings per system, the top five phrases were chosen, which amounts to 30 phrases per image. However, because of a few phrases overlapping between the systems, image “Proserpine” was accompanied by 27 distinct phrases to evaluate, image “Loss of Schooner” had 28 phrases, while image “Sunlight” had 29. For image “House in Provence”, human-generated phrases were also included. In total, participants were asked to evaluate 121 distinct phrases. According to some participants, the experiment required 15 minutes to complete.

Interface

The experiment was conducted through an online interface, with minimal changes compared to the final pilot experiment (§8.4.3). The only notable changes were that **i**) tags previously presented to participants as “what Mary thought” are now omitted and **ii**) underlining of end-words (tags) in the phrases was eliminated, in order to reduce the participants’ cognitive load, given the high number of phrases to be rated. The order between images was differentiated (as in the first and the third pilot) while the order of the phrases was simply randomised. Each image was presented on a different page, along with a list of phrases to be assessed. Each phrase could have been presented on a separate page, however, a simpler presentation was preferred for practical implementation reasons.

After assessors had been asked to evaluate the given phrases, they were asked to provide their own phrases. The last request was meant to encourage participants to produce phrases that they would expect to see in this experiment. Possible regularities in these phrases (e.g. particular phrases being suggested by most participants) would indicate what phrases should have been retrieved in higher ranks (high enough to be included in the experiment). Such phrases cannot be used as a surrogate for recall, since they would be too limited to provide a representative sample of the potentially vast number of phrases that could be considered acceptable for a given image. Additionally, phrases suggested by participants would tend to be disjoint from those used in the experiment (i.e. the ones in the highest ranks), hence, even if some notion of recall were to be measured, it would be uninformative, if not misleading. A screenshot from the interface can be seen in Figure B.5 (page 193).

Outcome

The evaluation experiment was completed by nine participants. The responses of each participant for the output of each system can be seen in Section B.4.2 (p. 194). The rest of this chapter provides details of the results obtained.

8.5.2 Inter-rater reliability

In the final pilot experiment, evaluators providing acceptability scores for phrases on a 5-point scale reached a moderate agreement, possibly because the task is difficult to perform. A similar level of agreement was expected in the main experiment, which used the same scoring system. Given the modest expectations, I decided to measure IRR in a more detailed fashion for the purpose of clarifying the nature of the consensus. Stemler (2004) has advocated the use of more thorough IRR estimates, arguing that seeing reliability as a single number⁸ is “at best imprecise, and at worst potentially misleading”. According to the author, IRR can be better described using three estimates: **i)** *consensus* estimates, which indicate the extent to which different judges agree with each other above what is expected by chance, **ii)** *internal consistency* estimates, which show the extent to which the responses of one rater can predict (i.e. correlate with) the responses of another rater even when their actual consensus is low and **iii)** *measurement* estimates, which quantify the distance between different raters with respect to their scoring behaviour. For this experiment, IRR will be reported on all three estimates.

Consensus estimate Consensus between participants was measured using Krippendorff’s α , as in the final pilot experiment. For the four images, IRR was moderate, as expected:

- “House in Provence” ($\alpha = 0.43$)
- “Proserpine” ($\alpha = 0.33$)
- “Loss of Schooner” ($\alpha = 0.35$)
- “Sunlight” ($\alpha = 0.34$)

To examine whether particular participants diverged from the rest, I also calculated IRR between each pair of participants, as shown in Table 8.1.

⁸The number ranges from -1 to 1 , where 0 means no consensus and 1 means perfect consensus and -1 means systematic disagreement.

Table 8.1: Pairwise Krippendorff’s α between participants

	P0	P1	P2	P3	P4	P5	P6	P7	P8
P0		0.14	0.49	0.56	0.52	0.19	0.51	0.40	0.28
P1			0.31	0.15	0.23	0.42	0.28	-0.15	0.51
P2				0.45	0.61	0.22	0.52	0.29	0.40
P3					0.50	0.30	0.52	0.28	0.26
P4						0.28	0.51	0.34	0.44
P5							0.36	-0.01	0.63
P6								0.32	0.39
P7									0.06
P8									

On the table, it can be observed that two participants ($P5$ and $P8$) are similar to each other (agreement 0.63) but very different from most other participants. Another participant ($P1$) seemed to be closer to $P5$ and $P8$ than the other participants. Thus, it is likely that the participants may come from (at least) two different populations, reflecting different phrase scoring patterns. Overall consensus between the nine participants is 0.37, however, if group $P1, P5, P8$ is separated from group $P0, P2, P3, P4, P6, P7$, agreement is 0.46 and 0.53 respectively.

Internal consistency estimate Internal consistency of raters was measured using Spearman’s ρ correlation coefficient (Spearman, 1904). Each participant was represented as a vector (called *scoring vector*) of length 121, corresponding to the 121 phrases that each one had been asked to rate. Correlation between the scoring vectors of pairs of participants can be seen in Table 8.2. The table shows that, even in cases where the pairwise consensus is low (see Table 8.1), correlation remains roughly at 0.5, which means that some of the variation might be systematic. A pair of raters whose responses have moderate correlation, that is consistency, but small α values, that is consensus, might have similar attitudes towards individual images, yet one is a lower rater than the other. For instance, the scoring vectors of $P0$ and $P1$ have $\rho = 0.5$, but $\alpha = 0.14$, suggesting that they have different levels of severity. Indeed, the average score of $P0$ for all phrases from both POS-Dep and Ngram is 2.98 while the average score of $P1$ is only 1.61.

Table 8.2: Pairwise Spearman ρ correlations between participants

	P0	P1	P2	P3	P4	P5	P6	P7	P8
P0		0.5	0.52	0.39	0.53	0.46	0.52	0.56	0.5
P1			0.47	0.25	0.44	0.49	0.57	0.43	0.51
P2				0.38	0.58	0.32	0.53	0.45	0.48
P3					0.41	0.28	0.39	0.34	0.31
P4						0.42	0.49	0.51	0.52
P5							0.54	0.49	0.63
P6								0.5	0.47
P7									0.36
P8									

Measurement estimate The next step is to provide a measurement estimate, that is represent each participant’s scoring behaviour by a summary number, which allows for easy comparison between the different respondents’ severity. Judge severity could easily be measured by the average score of a participant provided for all phrases for both systems, however, a simple average score does not account for variance. A commonly used method for measurement estimates is Principal Component Analysis (PCA) (Harman, 1967). If participants are represented as scoring vectors of length 121, then a single summary number can be given if the 121 dimensions are collapsed into one. Participant scores previously represented in a 121-dimensional vector space can now be single points on a line.⁹

Dimensionality reduction helps eliminate some of the noise in the data, which can then be represented in a simpler, more interpretable form. With PCA, I reduced participants’ ‘scoring vectors’ from 121 dimensions to one, as suggested by Stemler (2004). The single dimension to which all dimensions were collapsed captured 52% of the original variance, so it allows for a fair, though not perfect, representation of the multi-dimensional data. Participants’ ‘loadings’ (i.e. positions) in the first dimension have been plotted below:



Figure 8.5: Participants’ scoring vectors reduced to one dimension

As can be seen, participants P_5 and P_1 are low raters, which explains why the consensus estimate between them is high while their degree of agreement with the rest of assessors is low.

8.5.3 System performance

In the main evaluation experiment, participants were asked to assess phrase quality on a 5-point scale. A given phrase is not marked as ‘acceptable’ or ‘non-acceptable’; instead, it has a score attached to it indicating the phrase’s degree of acceptability. The score for each phrase, henceforth *phrase score*, is equal to the mean value of its ‘response vector’ (see §8.4.3), that is the average score assigned to it by the assessors. The score for each system is obtained by averaging all phrase scores associated with the system.

Graded, as opposed to binary, judgements imply that metrics such as precision in the widely understood sense ($TP/(TP + FP)$) can no longer be used. However, the score for each system can be seen as a surrogate for precision, in that it indicates how correct (i.e. acceptable) the output is.

Each one of the two systems in the evaluation experiment (POS-Dep and Ngram) was represented as a set of phrase scores. These two sets were treated as samples whose means (overall system scores) could be compared for statistical significance. One consideration was that, due to the difficulty of the acceptability judgement task, assessors had difficulty

⁹Projection to a plane (two dimensions) or hyperplane (higher dimensions) is also possible.

reaching consensus on the score of particular phrases. Thus, the response vector of some phrases (i.e. the vector of raw scores a phrase received from each participant) may have high variance. This information is lost if only the phrase score (mean value of the vector) enters the significance test. Another issue is that, with respect to IRR, there seemed to be two groups of participants that reached more consensus within them than between them (see §8.5.2), implying that there should be two distinct scorings of the systems, one for each of the two participant groups. To perform hypothesis testing while accommodating the above concerns, I applied three significance tests: **i)** a basic test, comparing the overall score of POS-Dep with that of Ngram, **ii)** a test that accounts for the two different consensus groups in the experiment and **iii)** a test that accommodates the variance within response vectors.

Basic hypothesis testing

The POS-Dep system had an average score of 2.72 while the Ngram system had an average score of 2.12. POS-Dep performs better with a difference of 0.6. Using the non-parametric test described in Section 4.2.2, I found that the *POS-Dep* system performs better than the *Ngram* system at a 99.2% level.

Hypothesis testing with different consensus groups

To account for the existence of different consensus groups, I split the respondents into the two groups that seemed to emerge (see §8.5.2) and compared the performance of POS-Dep and Ngram based on the responses of each group separately. The results of the significance tests are as follows

- $\{P0, P2, P3, P4, P6, P7\}$ → difference between POS-Dep and Ngram significant at 99.8% confidence level
- $\{P1, P5, P8\}$ → difference between POS-Dep and Ngram significant at 99.9%

Hypothesis testing with different variances within response vectors

Since IRR is modest in this experiment, it follows that there is a fair amount of variance in the different responses given by participants on every phrase. Phrases for which assessors have more trouble reaching a consensus are associated with high uncertainty with respect to their true score. It would be interesting to conduct a statistical test that corrects, or accounts for, this uncertainty.

For this statistical testing, the hypothesis is that phrases which divide opinions (i.e. have a lot of variance in their response vector) contain more *measurement error* than phrases with a general consensus. In the field of metrology, different measurements of a single entity or phenomenon (e.g. $\text{measurement1} = \{1.23, 1.20, 1.24, 1.23, 1.19\}$, $\text{measurement2} = \{\dots\}$ etc.), which may have been made by different machines, in different research groups or at different times, can be averaged in a way that has been proven to reduce the overall variance – and, by extension, uncertainty – of the measurement (Meier, 1953). The process is to produce a variance-weighted mean of the different measurements in a way that the contribution of a particular measurement to the grand mean is inversely proportional to its variance. For an introduction to different versions of this technique see (Bevington and Robinson, 1969) and (Taylor 1997; chapter 7).

In this experiment, the group of nine ratings provided for a given phrase by different participants can be seen as a measurement. It is important to clarify that ratings of different phrases are not measurements of the same thing, therefore, the variance-weighted average cannot be *proven* to reduce the overall uncertainty, but it can help inform a statistical conclusion on the basis of phrase scores that participants are more confident about. For this test, I have used the most common variant of uncertainty-weighted average, the Graybill-Deal estimator (Graybill and Deal, 1959). The notation is taken from (Zhang, 2006).

We can imagine a matrix with i rows and j columns, where each row contains the different ratings (nine in this experiment; one for each rater) that represent a phrase. Each column represents the responses of a particular participant. The mean score of each phrase is given by:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \quad (8.10)$$

where \bar{X}_i is the mean of phrase i , n_i is the number of ratings for phrase i . In this experiment, all phrases have exactly nine measurements. The weighted mean is:

$$\bar{X}_{GD} = \sum_{j=1}^k \hat{w}_i \bar{X}_i \quad (8.11)$$

where \bar{X}_{GD} stands for graded average, k is the number of phrases (i.e. number of rows) and \hat{w}_i is the weight for the score \bar{X}_i of a particular phrase, when it is used to computer the grand mean. The weight is given by:

$$w_i = \frac{\frac{1}{\sigma_i'^2}}{\sum_{j=1}^k \frac{1}{\sigma_j'^2}} \quad (8.12)$$

where $\sigma'^2 = \frac{\sigma^2}{n_i}$. Since population variances are usually not known, they can be estimated by means of sample variances:

$$\hat{w}_i = \frac{\frac{1}{s'^2}}{\sum_{j=1}^k \frac{1}{s'^2}} \quad (8.13)$$

The above equation assumes that there are no zero variances, which, in metrology is a reasonable assumption, since many different individual measurements are taken and with sensitive tools, so it is unlikely that a quantity has been measured, say, ten times with all decimal units identical. In this experiment, however, there were a few examples of phrases with zero variance (e.g. some phrases voted with ‘1’ by all nine participants). To correct for this, I randomly chose a participant and changed their measurement to ‘2’, which introduced only little additional variance.

After phrase scores of the baseline and POS-Dep were averaged as shown above, the significance test found that POS-Dep outperforms the baseline, at a confidence level of 99.8%.

Comparison with human-generated phrases

As mentioned in Section 8.5.1, next to image “House in Provence”, participants were presented with three, instead of two, types of output: **i)** phrases from POS-Dep, **ii)** phrases from Ngram and **iii)** phrases from unseen parallel subcorpus (henceforth *human phrases*), in which humans had described this particular image. The last set of phrases are essentially human generated, since they are extracted from image-specific text. The reason for the inclusion of human phrases for this image was two-fold:

1. **measuring acceptability:** If human phrases receive high ratings from assessors, this will demonstrate that phrases extracted from IDPs of the parallel subcorpus are indeed of high quality. This, in turn, will legitimise the use of the parallel subcorpus as reference for *a priori* judgements in the initial experiment (see §8.3.1).
2. **comparing with POS-Dep:** If the acceptability scores of human phrases are not significantly different from those of POS-Dep, this would imply that evaluators cannot distinguish the phrases automatically generated by the POS-Dep system from the, essentially, human-generated phrases of the parallel subcorpus.

The mean score obtained for human phrases in the experiment was 3.2, with individual phrase scores ranging from 2.3 (“grey mountains”) to 4.0 (“house in countryside”). This score was found to be significantly higher than that of POS-Dep for the same image (2.2), at a 99.7% confidence level. It was also, unsurprisingly, significantly higher than that of Ngram (1.83) with 99.93% confidence. As can be seen, even though assessors find the output of POS-Dep more acceptable than that of the baseline, they still find the phrases from the parallel data more acceptable than those of POS-Dep. If nothing else, such a result indicates that human phrases were of high acceptability, which justifies the use of the parallel subcorpus as a testbed for the initial experiment (§8.3.1). This does not give the subcorpus a gold standard status, given its size, but it provides reassurance that the phrases against which POS-Dep and Ngram were compared are highly acceptable.

8.5.4 Phrases suggested by participants

After assessing the phrases generated by the two systems for each image, participants were asked to suggest their own phrases. The purpose of this part of the experiment was **i)** to see whether participants could agree on a set of phrases that are essential for the given image and **ii)** if such phrases exist, examine whether the POS-Dep system returns them in a high enough rank.

Participants were advised to submit at least three phrases, from two to four words each, but were not prompted if phrases of other lengths or fewer phrases were provided. The reason was to avoid forcing participants to devise phrases that they did not consider necessary for the image, or phrases that they find unnatural because of the length restrictions. In total, 100 phrases were submitted; for each image, seven participants provided exactly three phrases, one participant provided four and one provided none. Most phrases submitted were two to four words long, as suggested, although larger phrases and single

words were also provided. The exact phrases submitted by the participants can be seen in the left column of Table B.5 (page 198).

In order to make sensible comparisons between the user-generated phrases and those of POS-Dep, it was important that the raw phrases, as submitted, were converted to a form producible by POS-Dep. The raw phrases were normalised following the steps below in the specified order:

1. converting every character to lowercase (e.g. “Bright and sunny” → “bright and sunny”)
2. removing diacritic marks (e.g. accents: Cézanne → “cezanne”)
3. splitting phrases that cannot be captured by the dependency patterns of POS-Dep (see §7.2) into sub-phrases that can be captured (e.g. “house with few windows” → “house with windows” + “few windows”). Sub-phrases that cannot be captured are omitted (e.g. “white house in valley in front of mountains” → “white house” + “house in valley”; the rest is omitted)
4. omitting responses that are not phrase-like, but a simple concatenation of words (e.g. “House trees mountains distance”) since they are misunderstandings of the task. From such responses, any phrase-like parts were kept (e.g. “stormy rescue mission ocean” → “rescue mission”)
5. omitting one-word phrases (e.g. “countryside”)
6. omitting – among the so far normalised phrases – those whose end words are not *both* tags from the image’s tag cloud (e.g. “woman eating fruit” is preserved because “woman” and “fruit” are both tags in the image’s tag cloud, but “beauty in dress” is omitted because only “dress” appears as a tag)

Table 8.3: Normalised phrases recommended by evaluators

“House in Provence”	“Sunlight”
country house (<i>3 times</i>)	white dress (<i>5 times</i>)
house in countryside (<i>2 times</i>)	bright and sunny
white house (<i>2 times</i>)	hill overlooking lake
blue mountains (<i>2 times</i>)	lady in white
blue hills	lady on hill
blue sky	lake landscape
cezanne painting	white lake
french countryside	woman looking against sunlight
french house	woman looking at water
hills and sky	woman looking beyond lake
house by mountains	woman in dress
house in mountains	woman in landscape
landscape and house	woman in lake
mountains behind trees	woman on hill
painting landscape	woman on hillside
rural scene	
trees surrounding house	
“Proserpine”	“Loss of the Schooner”
blue dress (<i>5 times</i>)	boats distress at sea
woman in dress (<i>3 times</i>)	rescue from sinking
painting of woman (<i>2 times</i>)	shipwreck at sea
woman eating fruit (<i>2 times</i>)	ships in storm
beautiful woman	ships in sea
black hair	sinking boats
blue velvet	sinking due to storm
blue woman	sinking ships
rosetti woman	stormy sea
velvet dress	waves of sea
woman and fruit	
woman holding pomegranate	

Conversion of each submitted phrase to its normalised equivalent(s) can be seen in Section B.4.3 (page 197). The normalised phrases, sorted by frequency, can be seen in Table 8.3.

As can be seen in the above table, participants seemed to agree on very few phrases as important descriptors of the image (e.g. “white dress” for image “Sunlight”, which occurs five times after normalisation, “blue dress” for image “Proserpine” and “country house” for image “House in Provence”). Most normalised phrases submitted by the evaluators are hapax legomena (i.e. have been suggested only once), possibly due to the small size of the data, which prevents any distribution from forming. If these phrases are seen as a sample of an underlying population of all possible phrases that are acceptable for each image, then this sample cannot be representative, given that the population might be indefinitely large. In fact, it is hard to prove that even larger samples are representative of the population (i.e. contain minimal sampling error), since the size of the latter is not

known. Even if it is assumed that the population of possible phrases follows a power-law distribution (where a small number of phrases can be enough to capture most probability mass), hapaxes in the above sample provide no information as to what proportion of the population they represent. In other words, it is not possible to determine whether a particular phrase that has been suggested only once by participants (e.g. “woman and fruit”) has been sampled from the top or from the long tail of the underlying population’s distribution. What this demonstrates is that attempting to estimate recall based on the above data would not be informative.

Phrases that were not hapaxes after normalisation (e.g. “painting of woman”, appearing twice) may potentially be seen as phrases on which there is some agreement among the participants. It would be interesting to see whether POS-Dep has retrieved them. For image “Proserpine”, the phrase “blue dress” (appearing five times in the normalised submitted phrases) had been returned by POS-Dep in rank 1, “woman in dress” (appearing three times) was in rank 2, “painting of woman” (appearing twice) was in rank 4, while “woman eating fruit” had been missed. For image “Sunlight”, the phrase “white dress” (appearing five times) had been returned by POS-Dep in rank 2. For image “House in Provence”, non-hapaxes are still produced, but in lower ranks: “country house” (appearing three times) was in rank 30 of POS-Dep, “blue mountains” (appearing twice) was in rank 39, “house in countryside” (appearing twice) was in rank 59, while “white house” was in rank 61. On the one hand, POS-Dep has produced most of the above non-hapaxes, which shows that the system was able to generate phrases which humans may suggest as important. On the other hand, the ranking was satisfactory in only half of the cases, which is probably due to the manual parameters initially assigned to the ranking function (§7.2.2). However, these parameters can be tuned, as suggested in Section 9.2.

8.6 Summary

In this chapter, I explained my evaluation decisions and described evaluation experiments for assessing the quality of the relation extraction system presented in the previous chapter. First, I justified the use of two testbeds, *a priori* and *a posteriori* human judgements, and motivated the choice of an ngram-based baseline. After presenting encouraging results on the first testbed, I discussed the challenges of designing an experiment to elicit *a posteriori* human judgements. Then, I described three pilots experiments and their outcomes. Finally, I provided details of the main evaluation experiment, in which the system proposed significantly outperformed the baseline.

Chapter 9

Conclusions

9.1 Contributions of the thesis

This thesis has approached collaborative tagging as a system that can capture and communicate the meaning of a document. Based on the intuition that this meaning might be delivered by tags working *together* rather than independently, the thesis sets out to investigate the existence of potential links between tags annotating particular documents. This exploration was based on a comparison between tags as they appear in tag clouds of images and natural language words as they appear in coherent text. The contributions of this thesis are summarised below.

A cross-disciplinary understanding of the image-tag relationship In Chapter 2, I examined the relationship between a tag and the image it annotates by bringing together theories from Library and Information Science (Otlet, 1934; Briet, 1951; Hutchins, 1978; Buckland, 1998; Shatford, 1986), Semiotics (Shannon and Weaver, 1964; Barthes, 1964; McLuhan, 1964), History of Art (Panofsky, 1955) and Information Retrieval (Salton and Harman, 2003; Voorhees and Harman, 2005). First, I explained that documents can be nested (e.g. a digital image of a painting of a book) and showed that tags that label an image can also label any of its enclosed documents. Then, I discussed the role of tags as subject indicators, demonstrating that they can act as *theme* (essential information about the image) or *rheme* (peripheral information). I also illustrated that tags can have an *of-ness* or *aboutness* relationship with the image; the former relating a tag with a concrete object depicted in the image and the latter relating a tag with an abstract concept expressed by the image. Finally, I utilised the theory of pre-iconography, iconography and iconology in order to show how tags can reveal different levels of understanding of the image.

Evidence that tags behave as words In Chapter 3, I showed that tags in a folksonomy follow a zipfian distribution (Zipf, 1932, 1949), as words do in a text corpus. I also found that that 97.6% of tag vocabulary items (types) were either single words or combinations of words from the Wikiwoods corpus (Flickinger et al., 2010). The vast majority of tag types (97.3%) were possible to find as nouns or adjectives in text, implying that tags are mostly used to describe entities from images and their attributes. In Chapter 4, I used distributional semantics to demonstrate that tags in tag clouds have similar syntagmatic and paradigmatic relations to words in text. This similarity was significantly

higher in the Steve folksonomy (§2.1) than in a semi-random version of Steve. Results of this chapter provided evidence that tags combine as words, which suggests that tag clouds are cohesive, and possibly coherent, entities.

A new parallel corpus In Chapter 5, I described the creation of a parallel corpus of tags and text provided for particular images. Such a corpus did not exist prior to this research and was crucial for investigating whether text submitted by users reveals how the same users’ tags are associated. After experimental design, which established variables, conditions, participant groups and stimuli, three pilot experiments were conducted. The final experiment collected tags and text from users with instructions based on real-life scenarios. The final corpus consists of 1,090 parallel tags-text annotations, provided by 218 participants. The corpus is publicly available.

Evidence for the existence of implicit inter-tag relations In Chapter 6, I used a subset (approximately 2/3) of the corpus described in Chapter 5 in order to study users’ tags in parallel with their text. After splitting multi-word tags and lemmatisation, I found that 55.6% of individual users’ tags appear in their own textual descriptions. When tags and descriptions were studied at a collective level (i.e. by all users simultaneously), the overlap was 73.8%. Another interesting finding was that 18.8% of a user’s possible tag pairs appear in the same sentence within their own descriptions. At a collective level, 39% of all possible tag pairs co-occur in sentences. Following this, I found that approximately 1/3 of unlemmatised tag pairs found in text were connected with continuous and unbranched dependency paths, providing evidence for the existence of inter-tag relations. Dependency patterns connecting tag pairs were extracted so that they could be used to learn further relations in the future.

Methods for suggesting image-specific relations without image-specific text In Chapter 7, I explained the motivation and design of a proof-of-concept system which aimed to show that it is possible to postulate well-formed and plausible tag-relation-tag triples for particular images using a text corpus alone. Since text accompanying tagged images is hard to encounter, the system had to rely on a generic corpus, such as Wikipedia, to find the desired relations. Visual processing was not performed, out of theoretical interest to examine the extent to which the problem is solvable with linguistic means alone. In Chapter 8, I discussed evaluation decisions that I made in order to measure the quality of the tag-relation-tag triples suggested from the system. On both evaluation testbeds (unseen data from parallel corpus and human judgements), the system outperformed the baseline.

9.2 Further work

Semantic interpretation of of Instantiated Dependency Patterns In this thesis, tag-relation-tag triples have been represented as dependency paths (catenae), such as *picture* \xleftarrow{ncsubj} *looks* \xrightarrow{ncmod} *like* \xrightarrow{dobj} *poster*, where “picture” and “poster” are tags of the same image. This representation makes the relationship between the two tags more explicit, however, it lacks the expressive power of a real semantic representation, such as one based on a predicate calculus. Future research could focus on converting

Instantiated Dependency Patterns (IDPs) to semantic representations by exploring issues such as **i**) lexical semantics, **ii**) quantification and **iii**) structural ambiguities. Lexical semantics will be needed to map a tag to a sense. For instance, ambiguous tags (e.g. “race”) will require disambiguation while spelling variants (e.g. “can-can”, “can can”, “cancan”) and synonyms (e.g. “automobile” and “car”) will need to be collapsed to one sense. A tag can be disambiguated in the context of its tag cloud using standard Word Sense Disambiguation techniques (e.g. Yarowsky 1993). Synonyms or near-synonyms can be detected using lexicons or distributional similarity. To make decisions regarding quantification, it will be necessary to identify the tags whose senses will be quantified. For instance, in an image depicting a girl and tagged with “girl”, it would be legitimate to produce a quantified expression such as $\exists x[\textit{girl}(x)]$. However, in an image tagged with “happiness”, an expression such as $\exists x[\textit{happiness}(x)]$ seems questionable. A distinction between concrete and abstract nouns might be informative. Quantifiers for plural nouns (e.g. generalised quantifiers) should also be investigated. Regarding structural ambiguity, the choice of representation formalism will be crucial. An example could be an IDP such as $\textit{people} \xrightarrow{\textit{ncmod}} \textit{having} \xrightarrow{\textit{dobj}} \textit{cocktail}$, which could be interpreted as “some people are having a cocktail each” or “there is a cocktail and some people are sharing it”. Resolving this ambiguity in the context an image can be a hard task, therefore, it might be useful to utilise a semantics that allows for structural underspecification, such as Minimal Recursion Semantics (MRS) (Copestake et al., 2005). Using MRS can also facilitate creating partial representations (e.g. $\textit{blue}(e_1)$, $\textit{ARG1}(e_1, x)$, $\textit{yellow}(e_2)$, $\textit{ARG1}(e_2, x)$) from the IDP $\textit{blue} \xleftarrow{\textit{conj}} \textit{and} \xrightarrow{\textit{conj}} \textit{yellow}$). Converting IDPs to semantic representations can raise interesting issues such as a single dependency structure (e.g. $* \xrightarrow{\textit{ncmod}} * \xrightarrow{\textit{dobj}} *$) mapping to two different semantic structures (e.g. one for “view of sea” and one for “crowd of people”), or two different dependency structures (e.g. $* \xleftarrow{\textit{ncmod}} *$ and $* \xleftarrow{\textit{ncsubj}} * \xrightarrow{\textit{xcomp}} *$) mapping to a single semantic structure (e.g. for “tall person” and “person is tall”).

Merging Instantiated Dependency Patterns Another extension of this work would be merging the generated IDPs into more complex structures. For example, $\textit{white} \xleftarrow{\textit{ncmod}} \textit{house}$ and $\textit{house} \xrightarrow{\textit{ncmod}} \textit{in} \xrightarrow{\textit{dobj}} \textit{countryside}$ can be connected into $\textit{white} \xleftarrow{\textit{ncmod}} \textit{house} \xrightarrow{\textit{ncmod}} \textit{in} \xrightarrow{\textit{dobj}} \textit{countryside}$. Merging can also be performed at the semantic level. For instance, although the IDP $\textit{blue} \xleftarrow{\textit{conj}} \textit{and} \xrightarrow{\textit{conj}} \textit{yellow}$ is hard to connect with the IDP $\textit{blue} \xleftarrow{\textit{ncmod}} \textit{coat}$ in a single catena, their possible MRS representations $\textit{blue}(e_1)$, $\textit{ARG1}(e_1, x)$, $\textit{yellow}(e_2)$, $\textit{ARG1}(e_2, x)$ and $\textit{coat}(y)$, $\textit{blue}(e)$, $\textit{ARG1}(e, y)$ can be merged into $\textit{coat}(x)$, $\textit{blue}(e_1)$, $\textit{ARG1}(e_1, x)$, $\textit{yellow}(e_2)$, $\textit{ARG1}(e_2, x)$ after unification of $\textit{blue}(e_1)$ with $\textit{blue}(e)$ and unification of $\textit{ARG1}(e_1, x)$ with $\textit{ARG1}(e, y)$.

Iteratively improving system suggestions In Section 7.2.2, I presented the equation used for assigning a plausibility score p to IDPs with respect to the images they had been suggested for. The parameters of the equation were initially set through manual experimentation with different sets of images, however, it is possible to optimise them given human judgements. For example, during the main evaluation experiment, humans assigned scores to 120 phrases. These scores can be used to help tune the parameters a_i of the equation $p = f(c, s, t, b, m) = a_1c + a_2s + a_3t + a_4b + a_5m$, in order to improve future plausibility predictions. Using multiple linear regression with ordinary least squares

optimisation, I fitted a hyperplane to a 120-dimensional vector space, such that the five independent variables of the equation (c for ‘corpus probability’, s for ‘similarity’, t for ‘tag weight’, b for ‘base form weight’ and m for ‘multi-word tag weight’) were used together to predict the dependent variable p (plausibility score). The co-efficients of the equation were set to $a_1 = 21.22$, $a_2 = 2.41$, $a_3 = 7.17$, $a_4 = 0.0$ and $a_5 = 1.04$. The first co-efficient (for phrase probability in corpus) is high because the probabilities are low. Parameters like these can be used to iteratively improve the results of the system with further judgements obtained in the future.

This thesis is the beginning of an exploration of collaborative tagging as a process of describing digital resources. Having shown that tags can express meanings in tandem and not just individually, the thesis paves the way for further research into the semantics of tag clouds and into methods for making explicit the implicit relations between the tags.

Bibliography

- Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. (2014). Representing Topics Labels for Exploring Digital Libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL-14) International Conference on Theory and Practice of Digital Libraries (TPDL-2014)*.
- Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS-13)*, pages 13–22.
- Angeletou, S., Sabou, M., and Motta, E. (2008). Semantically enriching folksonomies with FLOR. In *Proceedings of 5th Annual European Semantic Web Conference (ESWC-08) Workshop on Collective Semantics: Collective Intelligence and the Semantic Web*, pages 65–80.
- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Aston, G. and Burnard, L. (1998). *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.
- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2670–2676.
- Barthes, R. (1964). Rhétorique de l’image. *Communications*, 4(1):40–51.
- Benz, D. and Hotho, A. (2007). Position Paper : Ontology Learning from Folksonomies. In Hinneburg, A., editor, *Proceedings of LWA (Lernen, Wissen, Adaption; Learning, Knowledge, Adaption) Conference, September 2007*, pages 109–112.
- Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-12)*, pages 995–1005.
- Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL-99)*, pages 57–64.
- Bevington, P. and Robinson, D. (1969). *Data reduction and error analysis for the physical sciences*. McGraw-Hill, New York.

- Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Bollen, D. and Halpin, H. (2009). The Role of Tag Suggestions in Folksonomies. *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HYPERTEXT-09)*, pages 361–362.
- Briet, S. (1951). *Qu'est-ce que la documentation?* Editions Documentaires Industrielles et Techniques, Paris.
- Brin, S. (1999). Extracting Patterns and Relations. In *Selected Papers from the International Workshop on the World Wide Web and Databases (WebDB-99)*, pages 172–183. Springer Heidelberg, Berlin.
- Briscoe, T. (2006). An introduction to tag sequence grammars and the RASP system parser. Technical Report 662, University of Cambridge, Cambridge.
- Briscoe, T. and Carroll, J. (1995). Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the 4th International Workshop on Parsing Technologies (IWPT-95)*, pages 48–58.
- Bry, F. and Wieser, C. (2012). Squaring and scripting the ESP game: Trimming a GWAP to deep semantics. In *Proceedings of the 4th Human Computation Conference (HCOMP-12)*, pages 183–192.
- Buckland, M. (1998). What is a “digital document”? *Document Numérique*, 2(2):221–30.
- Cattuto, C., Benz, D., and Hotho, A. (2008). Semantic grounding of tag relatedness in social bookmarking systems. *Proceedings of the 7th International Semantic Web Conference (ISWC-08)*, pages 1–16.
- Chun, S., Cherry, R., Hiwiler, D., Trant, J., and Wyman, B. (2006). Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums. In *Proceedings of the 10th Annual Meeting of the Museums and the Web Conference (MW-06)*.
- Clarke, A., Elsner, M., and Rohde, H. (2013). Where’s Wally: the influence of visual salience on referring expression generation. *Frontiers in Perception Science (Special Issue on Scene Understanding)*, 4(1):392.
- Copestake, A., Flickinger, D., Pollard, C., and Sag, I. a. (2005). Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(2-3):281–332.
- Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, pages 77–86.
- Cruse, D. A. (1986). *Meaning in Language*. Oxford University Press, Oxford.
- Crystal, D. (2001). *Language and the Internet*. Cambridge University Press, Cambridge.

- Crystal, D. (2009). *Txtng: The Gr8 Db8*. Oxford University Press, Oxford.
- Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically Motivated Large-Scale NLP with C & C and Boxer. In *Proceedings of the 45th Annual Meeting of Association for Computational Linguistics (ACL-07) Interactive Poster and Demonstrations Session*, pages 33–36.
- de Beaugrande, R. and Dressler, W. (1981). *Introduction to Text Linguistics*. Longman, London.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Etzioni, O., Fader, A., and Christensen, J. (2011). Open Information Extraction: The Second Generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 3–10.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *Proceedings of the 11th European Conference on Computer Vision (ECCV-10)*, pages 15–29.
- Fellbaum, C. D. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. In Firth, J. R., editor, *Studies of Linguistic Analysis*, pages 1–32. Philosophical Society, Oxford.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Flickinger, D., Oepen, S., and Ytrestøl, G. (2010). WikiWoods: Syntacto-Semantic Annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-10)*, pages 1665–1671.
- Francis, W. N. and Kucera, H. (1982). *Frequency analysis of English usage*. Houghton Mifflin Company, Boston, MA.
- Fujita, S. (2000). Reflections on “Aboutness” TREC-9 Evaluation Experiments at Just-system. *National Information Standards and Technology Special Publication: : The Ninth Text REtrieval Conference (TREC 9)*, pages 281–288.
- Graybill, F. A. and Deal, R. B. (1959). Combining unbiased estimators. *Biometrics*, 15(4):543–550.
- Guy, M. and Tonkin, E. (2006). Folksonomies: Tidying up Tags? In *D-Lib Magazine*, <http://www.dlib.org/dlib/january06/guy/01guy.html> Accessed 12-04-2011.
- Hahn, M. and Meurers, D. (2011). On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In *Proceedings of the International Conference on Dependency Linguistics (DEPLING-11)*, pages 310–317.

- Halliday, M. (2014). *An Introduction to Functional Grammar*. Routledge, London and New York.
- Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.
- Halpin, H. (2009). Social Meaning on the Web: From Wittgenstein to Search Engines. In *Proceedings of the 1st Web Science Conference (WEBSCI-09)*, volume 24.
- Halpin, H., Robu, V., Shepherd, H., and Hall, W. (2007). The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th International Conference on the World Wide Web (WWW-07)*, pages 211–220.
- Harman, H. H. (1967). *Modern factor analysis*. University of Chicago Press, Chigaco.
- Harris, Z. (1954). Distributional Structure. *Word: Journal of the Linguistic Circle of New York*, 10(23):146–162.
- Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL-04)*.
- Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., New York.
- Hearst, M. a. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th conference on Computational Linguistics (COLING-92)*, pages 23–28.
- Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Mouton, Gravenhage, The Netherlands.
- Herrera, J., Peñas, A., and Verdejo, F. (2006). Textual Entailment Recognition based on dependency analysis and WordNet. In *Lecture Notes in Artificial Intelligence: Revised Selected Papers from the Machine Learning Challenges Workshop, (MLCW-05)*, pages 231–239.
- Heymann, P. and Garcia-Molina, H. (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical report, Stanford University.
- Heymann, P. and Garcia-Molina, H. (2009). Contrasting Controlled Vocabulary and Tagging: Do Experts Choose the Right Names to Label the Wrong Things? In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM-09), Late Breaking Results Session*, pages 1–4.
- Hinkel, E. (2004). Rhetorical Features of Text: Cohesion and Coherence. In *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*. Lawrence Erlbaum Associates, New Jersey.
- Hodges, P. E., Payne, W. E., and Garrels, J. I. (1998). The yeast protein database (YPD): A curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 26(1):68–72.

- Hotho, A., Robert, J., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference (ESWC-06)*, pages 411–426.
- Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM-13)*, pages 465–474.
- Hutchins, W. (1978). The concept of ‘aboutness’ in subject indexing. *Aslib Journal of Information Management*, 30(5):172–181.
- Ingwersen, P. (1992). *Information retrieval interaction*. Taylor Graham Publishing, London.
- Ipeirotis, P., Tamir, D., and Kanth, P. (2010). Mechanical Turk: Now with 40.92% spam. <http://www.behind-the-enemy-lines.com/2010/12/mechanical-turk-now-with-4092-spam.html> Accessed 22-10-2012.
- Jakobson, R. (1941). *Child Language, Aphasia and Phonological Universals*. Mouton, The Hague.
- Johns, A. M. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly*, 20(2):247–265.
- Khan, M. U. G., Nawab, R. M. A., and Gotoh, Y. (2012). Natural Language Descriptions of Visual Scenes Corpus Generation and Analysis. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-12) Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation*, pages 38–47.
- Kilgarriff, A. (1995). BNC database and word frequency lists. <http://www.kilgarriff.co.uk/bnc-readme.html> Accessed 06-03-2011.
- Kim, J.-T. and Moldovan, D. (1993). Acquisition of semantic patterns for information extraction from corpora. *Proceedings of 9th IEEE Conference on Artificial Intelligence for Applications (CAIA-93)*, pages 171–176.
- Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-03)*, pages 423–430.
- Krause, M. G. (1988). Intellectual problems of indexing picture collections. *Audiovisual Librarian*, 14(2):73–81.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage, Beverly Hills, CA.
- Kruijff, C. J. M. (2006). Dependency Grammar. In Brown, K., editor, *Encyclopedia of Language & Linguistics*. Elsevier Science Ltd.
- Kucera, H. and Francis, W. N. (1979). A Standard Corpus of Present-Day Edited American English, for use with Digital Computers: Manual of Information. Technical report, Department of Linguistics, Brown University.

- Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., and Choi, Y. (2013). Generalizing Image Captions for Image-Text Parallel Corpus. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*, pages 790–796.
- Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11)*, pages 1536–1545.
- Lehmann, F. and Wille, R. (1995). A Triadic Approach to Formal Concept Analysis. In Ellis, G., Levinson, R., and Rich, W., editors, *Conceptual Structures: Applications, Implementation and Theory (Lecture Notes in Artificial Intelligence vol. 954)*, pages 32–43. Springer, New York.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics*, 2(1):196–168.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, 22(140):5–53.
- Lin, D. (1993). Principle-Based Parsing Without Overgeneration. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, pages 112–120.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational linguistics (COLING-98)*, pages 768–774.
- Lin, D. and Pantel, P. (2001). DIRT - Discovery of Inference Rules from Text. *Natural Language Engineering*, 7(4):323–328.
- Lin, H., Davis, J., and Zhou, Y. (2009). An Integrated Approach to Extracting Ontological Structures from Folksonomies. In *Proceedings of the 6th Annual European Semantic Web Conference (ESWC) 2009*, pages 654–668.
- Maala, M. Z., Delteil, A., and Azough, A. (2007). A Conversion Process from Flickr Tags to RDF. In Flejter, D. and Kowalkiewicz, M., editors, *Proceedings of the 10th International Conference on Business Information Systems (BIS-07) Workshop on Social Aspects of the Web*.
- Magatti, D., Calegari, S., Ciucci, D., and Stella, F. (2009). Automatic labeling of topics. In *Proceedings of the 9th International Conference on Intelligent Systems Design and Applications (ISDA-09)*, pages 1227–1232.
- Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In Wolman, B. and Nagel, E., editors, *Scientific Psychology*. Basic Books, New York.
- Marinchev, I. (2006). Practical Semantic Web - Tagging and Tag Clouds. *Cybernetics and Information Technologies*, 6(3):3–9.

- Maron, B. (1977). On Indexing, Retrieval and the Meaning of About. *Journal of the American Society for Information Science*, 28(1):38–43.
- Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. In *Computer Mediated Communication*, <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html> Accessed 09-11-2011.
- McLuhan, M. (1964). *Understanding media: The extensions of man*. McGraw-Hill, New York.
- Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9(1):59–73.
- Melamud, O., Dagan, I., Goldberger, J., and Szpektor, I. (2013). Using Lexical Expansion to Learn Inference Rules from Sparse Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*, pages 283–288.
- Merholz, P. (2004). Ethnoclassification and vernacular vocabularies. <http://www.peterme.com/archives/000387.html> Accessed 10-04-2012.
- Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC-05)*, pages 522–536.
- Minnen, G., Carroll, J., and Pearce, D. (2000). Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference (INLG-00)*, pages 201–208.
- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP-09)*, pages 1003–1011.
- Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–429.
- Mitton, R. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management*, 23(5):495–505.
- Nakov, P. I. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. PhD thesis, University of California, Berkeley.
- National Information Standards Organisation (2004). *Understanding metadata*. NISO Press, Bethesda, MD.
- Newman, D., Lau, J., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10)*, pages 100–108.
- Nivre, J. (2005). Dependency grammar and dependency parsing. Technical report, School of Mathematics and Systems Engineering, Vaxjo University, Sweden.

- Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York.
- Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th Neural Information Processing Systems Conference (NIPS-11)*.
- O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, 65(1):17–37.
- Otlet, P. (1934). *Traité de documentation*. Editiones Mundaneum, Brussels.
- Panofsky, E. (1955). *Meaning in the visual arts*. Doubleday Anchor, Garden City, New York.
- Pantel, P. and Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-04)*, pages 321–328.
- Peters, I. and Stock, W. (2007). Folksonomy and information retrieval. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology (ASIS&T-07)*.
- Pink, D. (2005). Year in Ideas: Folksonomy. In *New York Times*, <http://www.nytimes.com/2005/12/11/magazine/11ideas1-21.html> Accessed 26-07-2013.
- Recanati, F. (2004). *Literal Meaning*. Cambridge University Press, Cambridge.
- Riezler, S. and Maxwell, J. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05) Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*, pages 57–64.
- Riloff, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the 13th National Conference of Association for the Advancement of Artificial Intelligence (AAAI-96)*, pages 1044–1049.
- Roy, R. S., Reddy, M. D., Ganguly, N., and Choudhury, M. (2014). Understanding the Linguistic Structure and Evolution of Web Search Queries. In *Proceedings of the 10th International Conference on the Evolution of Language (EVOLANG-14)*, pages 286–293.
- Sahlgren, M. (2008). The distributional hypothesis. In *Rivista di Linguistica (Special Issue on Distributional models of the lexicon in linguistics and cognitive science)*, volume 20, pages 33–53.
- Salton, G. and Harman, D. (2003). Information retrieval. In Ralston, A., Reilly, E., and Hemmendinger, D., editors, *Encyclopedia of Computer Science*, pages 858–863. John Wiley and Sons, 4th edition.
- Salton, G., Yang, C., and Yu, C. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

- Saussure, F. (1916). *Cours de linguistique générale*. Payot, Paris.
- Schmitz, C., Hotho, A., Jäschke, R., and Stumme, G. (2006). Mining Association Rules in Folksonomies. In *Proceedings of the 10th International Federation of Classification Societies Conference (IFCS-06)*, pages 1–9.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.
- Shannon, C. E. and Weaver, W. (1964). *The mathematical theory of communication*, volume 14. University of Illinois Press, Urbana, IL.
- Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62.
- Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2nd ACM Conference on Recommender Systems (RECSYS-08)*, pages 259–266.
- Shinyama, Y. and Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-06)*, pages 304–311.
- Shirky, C. (2005). Ontology is Overrated. <http://www.shirky.com/writings/ontology-overrated.html> Accessed 29-10-2011.
- Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1):233–272.
- Sordo, M., Gouyon, F., and Sarmiento, L. (2010). A method for obtaining semantic facets of music tags. In *Proceedings of the 4th ACM Conference on Recommender Systems (RECSYS-10) Workshop on Music Recommendation and Discovery*.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.
- Specia, L. and Motta, E. (2007). Integrating Folksonomies with the Semantic Web. In *Proceedings of the 4th European Web Semantic Conference (ESWC-07)*, pages 624–639.
- Spiteri, L. (2007). Structure and form of folksonomy tags: The road to the public library catalogue. In *Webology*, <http://www.webology.org/2007/v4n2/a41.html> Accessed 17-08-2012.
- Stainton, R. (2006). *Words and Thoughts : Subsentences, Ellipsis, and the Philosophy of Language*. Oxford University Press, Oxford.
- Steels, L. (2006). Collaborative tagging as distributed cognition. *Pragmatics and Cognition*, 14(2):275–285.
- Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, 9(4):66–78.

- Steyvers, M. and Griffiths, T. (2007). Probabilistic Topic Models. In Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors, *Handbook of Latent Semantic Analysis*, volume 22, pages 1028–1040.
- Strohmaier, M., Körner, C., and Kern, R. (2010). Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWS-10)*.
- Stubbs, M. (2008). Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence. In Schiffrin, D., Tannen, D., and Hamilton, H., editors, *The Handbook of Discourse Analysis*, pages 304–320. Blackwell, Oxford.
- Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. MIT Press, Cambridge, Mass.
- Taylor, J. R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Mill Valley, CA, 2nd edition.
- Tennis, J. T. (2006). Social Tagging and the Next Steps for Indexing. In *Proceedings 17th Annual Meeting of the Association for Information Science and Technology (ASIS&T-06) Special Interest Group on Classification Research*.
- Trabelsi, C., Jrad, A. B., and Yahia, S. B. (2010). Bridging Folksonomies and Domain Ontologies: Getting Out Non-taxonomic Relations. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM-10)*, pages 369–379.
- Trant, J. (2009a). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, 10(1):1–42.
- Trant, J. (2009b). Tagging, Folksonomy and Art Museums: Results of steve.museums research. In *Archives Museum Informatics*, <http://www.museumsandtheweb.com/files/trantSteveResearchReport2008.pdf> Accessed 23-02-2012.
- Trant, J. and Wyman, B. (2006). Investigating social tagging and folksonomy in art museums with steve.museum. In *Proceedings of the 15th International Conference on the World Wide Web (WWW-06) Workshop on Collaborative Web Tagging*.
- Travis, C. (1997). Pragmatics. In Hale, B. and Wright, C., editors, *A Companion to the Philosophy of Language*, pages 87–107. Blackwell, Oxford.
- Travis, C. (2000). *Unshadowed thought*. Harvard University Press, Cambridge, MA.
- Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.
- Vander-Wal, T. (2005). Explaining and showing broad and narrow folksonomies. <http://www.vanderwal.net/random/entryse1.php?blog=1635> Accessed 12-10-2011.
- Vander-Wal, T. (2007). Folksonomy. <http://vanderwal.net/folksonomy.html> Accessed 12-03-2010.

- Viethen, J. and Dale, R. (2011). GRE3D7: A Corpus of Distinguishing Descriptions for Objects in Visual Scenes. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP-11) Workshop on Natural Language Generation and Evaluation*, pages 12–22.
- von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine*, 39(6):92–94.
- Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, Mass.
- Weller, K. (2007). Folksonomies and ontologies: two new players in indexing and knowledge representation. In Jezzard, H., editor, *Applying Web 2.0: Innovation, Impact and Implementation. Proceedings of the Online Information Conference*, pages 108–115.
- Wieser, C., Bry, F., Alexandre, B., Lagrange, R. (2013). ARTigo: Building an Artwork Search Engine With Games and Higher-Order Latent Semantic Analysis. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP-13)*, pages 15–20.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishers, London.
- Wu, F. and Weld, D. (2010). Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 118–127.
- Yarowsky, D. (1993). One Sense Per Collocation. *Proceedings of the ARPA Human Language Technology Workshop (HLT-93)*, pages 266–271.
- Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 947–953.
- Yoon, J. and O’Connor, B. (2010). Engineering an image-browsing environment: repurposing existing denotative descriptors. *Journal of Documentation*, 66(5):750–774.
- Yule, G. (1912). On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.
- Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3(1):1083–1106.
- Zhang, N. (2006). The uncertainty associated with the weighted mean of measurement data. *Metrologia*, 43(3):195–204.
- Zhou, G., Su, J., and Zhang, J. (2005). Exploring Various Knowledge in Feature-based Relation Extraction. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-05)*, pages 427–434.
- Zipf, G. K. (1932). *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.
- Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, Mass.

Zubiaga, A., Martínez, R., and Fresno, V. (2009). Getting the Most Out of Social Annotations for Web Page Classification. In *Proceedings of the 9th ACM Symposium on Document Engineering (DOCENG-09)*, pages 74–83, Munich, Germany.

Appendix A

Parallel corpus experiments

A.1 Pilot 1

Responses

Below are the responses of the three participants in the first pilot experiment. Responses have been transcribed from paper, with spelling and other errors left intact.

Table A.1: Data from Pilot 1, Participant 1

	TAGS	DESCRIPTIONS
<i>moulin</i>	<ul style="list-style-type: none">-Silhouettes-Poster-Shadows-Offset Colour-Text-Dancing-Foreign Language-Charicatured	A stylised poster for Moulin Rouge. French text against a lively silhouetted background with a dancing girl and smoking gentleman in the foreground.
<i>torso</i>	<ul style="list-style-type: none">-Tribal-Psychedelic-Thick brush strokes	A black - possibly African? - lady stands in a traditional pose in the nude. Painted with broad strokes and natural greens and browns.
<i>angel</i>	<ul style="list-style-type: none">-Stained glass windows-Biblical-Ecclesiastical-Angelic-Saintly-Triumphalist-Last Trumpet-Resurrection	Elaborate stain-glassed religious artwork rapture and last trumpet.
<i>house</i>	<ul style="list-style-type: none">-House-Landscape-Trees-Grassland	Bland but scenic landscape. Moderate greenery with barren trees.

<i>sisters</i>	<ul style="list-style-type: none"> -Satanic -Girls -Children -Grainy 	Grainy artwork in which Satanic girl plots murder of angelic sister.
----------------	--	--

Table A.2: Data from Pilot 1, Participant 2

	TAGS	DESCRIPTIONS
<i>moulin</i>	art → poster → Paris → woman dancing	Moulin Rouge Poster - woman dancer - 19th century Paris
<i>torso</i>	art → abstract → tribal	A modern, abstract painting with a coloured person in the centre. Somewhat tribal.
<i>angel</i>	art → church art → angel → mosaic → stained glass windows	A stained glass window with an angel in the centre, most likely church art.
<i>house</i>	art → house	A house set against the mountainside surrounded by trees and shrubbery.
<i>sisters</i>	art → girls in red → creepy children	Two girls in red dresses against a blue wall with the taller girl having creepy eyes and being right out of children of the corn.

Table A.3: Data from Pilot 1, Participant 3

	TAGS	DESCRIPTIONS
<i>moulin</i>	<ul style="list-style-type: none"> -advertisement -Moulin Rouge -dancing -ball -silhouettes -man with top hat -dancing firl with wide skirt -La Goulue 	The advertisement shown is meant to attract people to some kind of ball in the style of Moulin Rouge which is happening every evening, presumably at a place called "La Goule". The picture itself shows a Moulin Rouge-style dancing girl in a kind of ballroom surrounded by some black silhouettes of people, male and female. The men are shown wearing to hars as is the one in the front of picture. He holds his right hand to his mouth and has his back hunched backwards. He is depicted in a brown colour.
<i>torso</i>	<ul style="list-style-type: none"> -person without a clear gender, maybe female -weird smile -disturbing colours -melancholic -modern art -abstract 	The picture, which is some sort of abstract art, shows a person standing slightly to the left of the middle. It is not quite clear of which gender the person is, but it may be a woman. The person is shown from the hips upwards, its left arm resting on the hops. The colours abscur the picture and you cannot really see whether ther is anything else in the backround.

<i>angel</i>	<ul style="list-style-type: none"> -church windows -ressurrection -archangel Michael -dominating colours blue, yellow and white -dedicated to Benjamin Harrison by his wife Mary -strange body armour of the archangel -trumpet to call out the last day 	<p>This picture shows an arrangement of church window. The theological main theme is resurrection. The middle window and the two ones on either side are mainly occupied by the archangel Michael who is depicted in a strange body armour under some clothes which resemble a Roman toga. He also has huge wings. His right hand points upward with an open palm, his left hand holds a trumpet which is said to use for the judgement day. Apart from the decoration the windows also show the sentence "Awake thou that sleepest. Arise from the dead and Christ shall give thee light." A sign below the middle window indicates that the window arrangement was dedicated to Benjamin Harrison by his wife Mary. The dominating colours of this early nineteenth century windows are blue, yellow and white.</p>
<i>house</i>	<ul style="list-style-type: none"> -mountains -house in the countryside -some dead trees -contrast between green and brown/beige colours at the bottom of the picture and white, grey and blue at the top. -neo-impressionism 	<p>The picture shows a two-floor-house in the countryside in front of some mountain formation. The house is set in a dirty white colour and is surrounded by trees, some of which are dead. You also can see some lawns, empty acres and a small round in the vicinity of the house. The mountains behind the house which occupy about 2/5 of the picture seem to have to levels and to be partly covered in snow. The picture seems to be painted in a sort of neo-impressionism.</p>
<i>sisters</i>	<ul style="list-style-type: none"> -art -pointilism -girls in red dresses -flower bouquet -blue background with arched brown frame 	<p>This picture shows two girls in a rather sterile, cold setting with a blue background and an arched brown frame. The taller and older one of the two girls is standing half way behind the smaller one resting her right arm on a table with a colourful tablecloth. Both girls wear the same red dress. They both have long hair. While the smaller girl in front looks at the viewer, the other one turns her eyes and head to the left (from viewer's perspective). On the table there is also a white flower bouquet. The style seems like a sort of pointilism.</p>

A.2 Pilot 2

Responses

Table A.4: Pilot 2, Participant 1 [copy from electronically submitted data]

	TAGS
<i>moulin</i>	<ul style="list-style-type: none"> -Moulin Rouge -Poster -Dancing Girl -Crowd -Top Hat Man
<i>african</i>	<ul style="list-style-type: none"> -Impressionist -Jungle -Woman -African -Body -Naked
<i>angel</i>	<ul style="list-style-type: none"> -Stained Glass -Angel Gabrielle -Blue Angel -Christ's light -The trumpeting angel
<i>house</i>	<ul style="list-style-type: none"> -House in Mountain -landscape -blue mountain
<i>sisters</i>	<ul style="list-style-type: none"> -Sisters -Posh -Portrait -Children

Table A.5: Data from Pilot 2, Participant 2 [copy from electronically submitted data]

	TAGS
<i>moulin</i>	<ul style="list-style-type: none"> -Moulin Rouge (La Goulue) -Poster -Twentieth century -Beige and red -Cabaret
<i>african</i>	<ul style="list-style-type: none"> -Oil painting -Female form -Modern art
<i>angel</i>	<ul style="list-style-type: none"> -Stained glass -Angel -Blue -Religious art -Awake thou that sleepest

<i>house</i>	<ul style="list-style-type: none"> -House on hill -Mountains -Landscape -Impressionism -Modern art -Trees
<i>sisters</i>	<ul style="list-style-type: none"> -Two girls -Framed painting -Twentieth century -Traditional interior -Blue background -Period dress

Table A.6: Data from Pilot 2, Participant 3

	DESCRIPTIONS
<i>moulin</i>	The picture is a pop art poster advertising Moulin Rouge. It shows a woman dancing and a man in the foreground. There are people watching in the background. The colors of the poster are red, white, yellow, brown and black.
<i>african</i>	The picture is an impressionistic painting that shows a naked woman. The colors that are used are rather limited: white, red, yellow, brown and black.
<i>angel</i>	The picture shows a window of a church or cathedral. It focuses on an angel in the middle of the picture. The colours are bright and a lot of blue is used.
<i>house</i>	The picture shows a house surrounded by a lot of nature and trees. There are mountains in the background.
<i>sisters</i>	The picture shows two girls dressed in formal clothing sitting at a table. There is a vase with flowers on the table. The picture has a blue frame.

A.3 Pilot 3

Responses

Table A.7: Data from Pilot 3, Participant 1

	PERSONAL INFORMATION
	Native speaker of English?: No Experience with tagging?: Yes
	DESCRIPTIONS

<i>moulin</i>	This is publicity add by Toulouse Lautrec, made to advertise the famous Parisian night club of Moulin Rouge. There is female dancer in the middle of the stage. Her image, in full color, contrasts with that of the audience, made up by black silhouettes.
<i>african</i>	This is a picture of a dark complexion male or female, possibly from Africa. He/she is naked. The painting is done with very thick strokes and is intentionally not well defined. The background consists mostly of white and other colors similar to the ones used for the human representation, tending to create an assimilation between them.
<i>angel</i>	This pictures shows a glass stained window of Gabriel the archangel calling the dead to arise from their tomb. This is a scene taken from the Book of Revelations, that represents the reckoning. Blue and gold are the dominant colors.
<i>house</i>	The landscape made with watercolor, shows a house in front of a desolated mountain range. The bleak grey color of the mountain establishes a contrast between the variety of colors that surround the landscape around the house.
<i>sisters</i>	The picture shows two Caucasian girls, probably sisters. They're standing next to a mantelpiece with a flower vase on top of it. They're also wearing red dresses. The painting is done with a technique of very small dots to convey the overall image.

Table A.8: Data from Pilot 3, Participant 2

	PERSONAL INFORMATION
	Native speaker of English?: Yes Experience with tagging?: No
	TAGS
<i>moulin</i>	-poster -moulin rouge -French -vintage -yellow -dance -can-can
<i>african</i>	-naked -woman -red -yellow -smudges -abstract -messy

<i>angel</i>	-religious -church -stained glass -angel -window symbolism man
<i>house</i>	-landscape -mountains -house -green and brown -quiet
<i>sisters</i>	-pretty -girls -sad faces -matching dresses -mosaic -muted colours -life like

A.4 Final experiment

A.4.1 Recruitment Email

"Please help me collect data for my experiment! (Chance to win a £100 John Lewis voucher)"

Dear all,

I am carrying out a fun and easy online image description experiment and I would be grateful if you could take a few minutes to participate. All you need to do is briefly describe some pictures of artworks. In exchange for this favour, you will enter a draw for a £100 John Lewis voucher.

****PROCESS****: The experiment is in two phases. Each phase will take you approximately 10 minutes to complete. (To get started with Phase 1, click on the link provided at the end of the email . After approximately two weeks, you will receive an email asking you to complete Phase 2.

Please note that a chance to win the £100 voucher:

- i) is only given to those completing **both** phases of the experiment
- ii) does not depend on a participant's responses

****DATES****:

Phase 1: 8-15 February

Phase 2: 1-8 March

The experiment has been approved by the Ethics Committee of the Computer Laboratory. Completion of the survey is voluntary and is not tied to any academic obligations.

To start Phase 1 of the experiment, please click on the link below:

<http://www-dyn3.cl.cam.ac.uk/~tt309/experiment-lent-2013/code/phase1/index.cgi>

Many thanks in advance!

All the best,
Theodosia

--

PhD student
Natural Language Processing and Information Group
Computer Laboratory
University of Cambridge

A.4.2 Differentiating image order

The order of images displayed to participants in the parallel corpus experiment was differentiated on the basis of edit distance. For all possible sequences (permutations) of the five images used in the experiment, the maximum edit distance was 5, but edit distance 4 was also allowed.

This method was chosen as an alternative to randomisation and intended to avoid sequences that are too similar to each other in case of low participation. For instance, if only 10 participants successfully completed the experiment, each task would consist of only 20 out of 120 possible sequences for both task, and since smaller samples can introduce larger sampling error, it is likely that these 20 sequences would not be different enough to be representative of the underlying possibilities. In practice, a small number of sequences some of which are too similar would introduce unwanted variables to our analysis.

The problem of finding the best set of permutations all of which have maximum edit distance from each other is computationally intractable, so an approximate solution was attempted. The process was as follows:

1. Levenshtein distance (Levenshtein, 1966) was computed for all $\binom{120}{2}$ (i.e. 7,140) possible pairs from the 120 permutations
2. Only pairs with maximum (5) and second maximum (4) distance were saved. This step reduced the set of pairs from 7,140 to 4,380.

3. Information from the list of 4,380 pairs was recorded in a hash table mapping a sequence to a set of sequences of ‘approved’ distance to it. The hash table ended up having 120 keys, same as all the possible sequences.
4. A random sequence A was selected from the 120 permutations and a random sequence B was selected from the set of sequences stored as values of A in the hash table.
5. Sequence A was written in a file, followed by sequence B .
6. Key A (along with its values) was removed from the hash table, to the sequence is not re-used.
7. Steps 4-6 were repeated until the hash table was empty (i.e. all 120 permutations had been written in a file).

With the above process, all possible sequences were pre-computed and written in a file, with priority given to those that were different enough from each other. Every time a participant hit the “I’m ready to tag” or “I’m ready to describe” button, a new sequence of images was selected from the file, so that after the first 120 participants had submitted their data, the file was fully read and sequences were being re-used in the same order.

A.4.3 Phase One interface

Tagging version

Image Description Experiment

Phase 1

Thank you for participating in this experiment. It researches the language that people use to describe pictures.

You will be shown **5 pictures** and asked to describe them by typing into a text window. The experiment should take roughly 10 minutes. You have the right to abandon it at any point.

Please answer the following questions: *Asterisk (*) indicates a mandatory question*

1. Are you a native speaker of English? *

Yes
 No

2. Have you ever performed tagging? *
(on websites like Flickr, Delicious, CiteULike, Bibsonomy, LibraryThing, LastFm etc.)

Yes
 No

3. What is your academic email address? *

Your academic email address is collected as a means for us to contact you in approximately two weeks for Phase 2 of the experiment and put you in a draw for a £100 John Lewis voucher. This is the only piece of personal information collected and will be destroyed after completion of the prize draw in late February.

4. Please acknowledge the following: *

I agree to be contacted for the second phase of the experiment and I understand that I need to complete both phases to enter the draw for a £100 John Lewis voucher.

Figure A.1: Phase 1 Welcome Page (tagging)

Imagine that there is an art website which contains images of artworks; let's call it **www.my-personal-gallery.com**. This website allows you to register, choose your favourite art images and create a personal collection.

In order to organise your images and be able to find them in the future, the website allows you to label them with keywords (**tags**).

Now you will be shown **5 pictures**. Please provide tags for each one of them. You are free to type in anything as a tag as long as it helps you retrieve the picture from your collection later.

Figure A.2: Phase 1 Scenario Page (tagging)

Description version

Image Description Experiment

Phase 1

Thank you for participating in this experiment. It researches the language that people use to describe pictures.

You will be shown **5 pictures** and asked to describe them by typing into a text window. The experiment should take roughly 10 minutes. You have the right to abandon it at any point.

Please answer the following questions: *Asterisk (*) indicates a mandatory question*

1. Are you a native speaker of English? *

Yes
 No

2. What is your academic email address? *

Your academic email address is collected as a means for us to contact you in approximately two weeks for Phase 2 of the experiment and put you in a draw for a £100 John Lewis voucher. This is the only piece of personal information collected and will be destroyed after completion of the prize draw in late February.

3. Please acknowledge the following: *

I agree to be contacted for the second phase of the experiment and I understand that I need to complete both phases to enter the draw for a £100 John Lewis voucher.

Figure A.9: Phase 1 Welcome Page (description)

Imagine that you are in a bookshop holding a book in your hands. The book contains art images.

Next to you there is a person with impaired vision and they are asking you to describe and explain to them what the pictures are about.

Now you will be shown **5 pictures**. Please describe them to this person.

Figure A.10: Phase 1 Scenario Page (description)

Image 1 of 5

Please describe this image to the person you met at the bookshop.



Please write at least 3 lines

Next

Figure A.11: Phase 1 Image 1 Page (description)

Image 2 of 5

Please describe this image to the person you met at the bookshop.



Please write at least 3 lines

Next

Figure A.12: Phase 1 Image 2 Page (description)

Image 3 of 5

Please describe this image to the person you met at the bookshop.



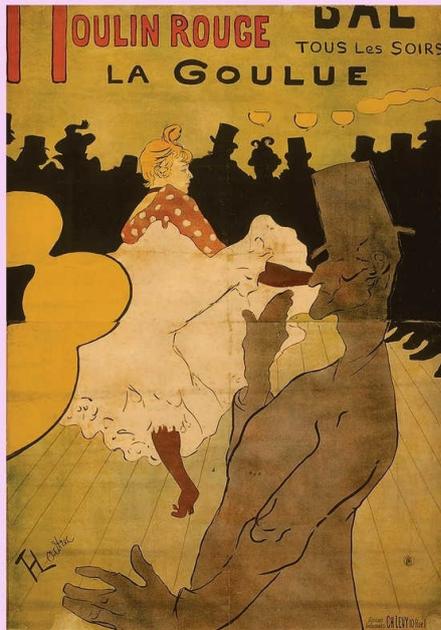
Please write at least 3 lines

Next

Figure A.13: Phase 1 Image 3 Page (description)

Image 4 of 5

Please describe this image to the person you met at the bookshop.



Please write at least 3 lines

Next

Figure A.14: Phase 1 Image 4 Page (description)

Image 5 of 5

Please describe this image to the person you met at the bookshop.



Please write at least 3 lines

Next

Figure A.15: Phase 1 Image 5 Page (description)

Thank you!

We are grateful that you participated in *Phase 1* of this experiment. You will be contacted in approximately two weeks in order to complete *Phase 2* and enter a draw for a **£100** John Lewis voucher.

Figure A.16: Phase 1 Thank You Page (description)

A.4.4 Phase Two interface

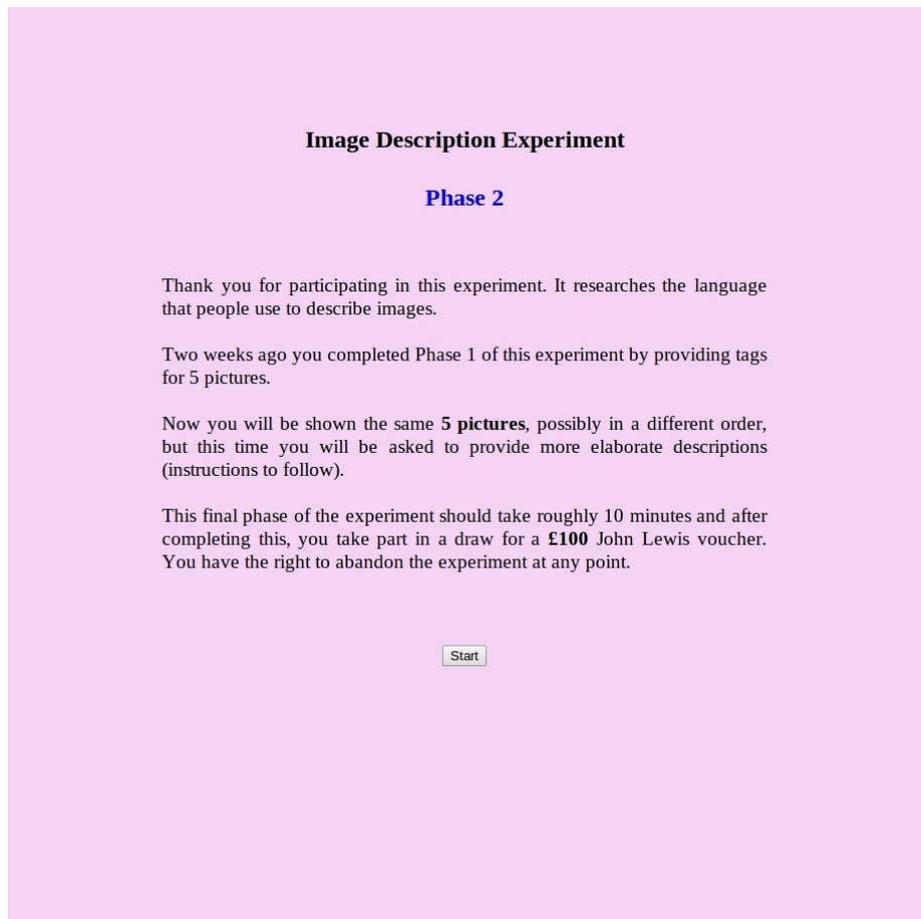


Figure A.17: Phase 2 Welcome Page (description)



Figure A.18: Phase 2 Thank You Page (description)

Image Description Experiment

Phase 2

Thank you for participating in this experiment. It researches the language that people use to describe images.

Two weeks ago you completed Phase 1 of this experiment by providing textual descriptions for 5 pictures.

Now you will be shown the same **5 pictures**, possibly in a different order, but this time you will be asked to provide much shorter descriptions (instructions to follow).

This final phase of the experiment should take roughly 10 minutes and after completing this, you take part in a draw for a **£100** John Lewis voucher. You have the right to abandon the experiment at any point.

Please answer the following question:

Have you ever performed tagging?

(on websites like Flickr, Delicious, CiteULike, Bibsonomy, LibraryThing, LastFm etc.)

- Yes
 No

Start

Figure A.19: Phase 2 Welcome Page (tagging)

Thank you!

You have now entered a draw for a **£100** John Lewis voucher. If you are successful, you will be notified by the end of March.

If you are interested in the results of the experiment, please contact Theodosia on t309@cam.ac.uk.

Figure A.20: Phase 2 Thank You Page (tagging)

Appendix B

Evaluation experiments

B.1 Pilot experiment 1 (with binary judgements)

B.1.1 Screenshots

Phrase Assessment Experiment

Thank you for volunteering! The experiment should take you at most 10 minutes.

You will be shown **6 pictures**. Next to each picture there will be a few phrases in note form (e.g. "house in countryside"). For each one of these phrases, please indicate if you think they are **good** or **bad** descriptors of the image.

An example is provided below.

Example



- "woman sitting on couch" → **Good** descriptor
- "wooden floor" → **Good** descriptor
- "porcelain on table" → **Good** descriptor
- "woman making carpet" → **Bad** descriptor (not relevant to the picture)
- "porcelain and to table" → **Bad** descriptor (non-sensical)

Figure B.1: Introductory page of first pilot evaluation experiment

Image 1 of 6

Image information

Title
<u>Bacchus and Ariadne</u>
Creator
Francois Boucher, 1703-1770
Created
1747-1766
Rights
2007 Museum Associates/LACMA
Institution
Los Angeles County Museum of Art



Please indicate if the phrases below are **good** or **bad** descriptors of the image.

	Good	Bad
mythical gods	<input type="radio"/>	<input type="radio"/>
battle gods	<input type="radio"/>	<input type="radio"/>
aubusson tapestry	<input type="radio"/>	<input type="radio"/>
dionysus bacchus	<input type="radio"/>	<input type="radio"/>
men were gods	<input type="radio"/>	<input type="radio"/>
men keep women	<input type="radio"/>	<input type="radio"/>
women than in men	<input type="radio"/>	<input type="radio"/>
battle against men	<input type="radio"/>	<input type="radio"/>
women are as men	<input type="radio"/>	<input type="radio"/>
women joined men	<input type="radio"/>	<input type="radio"/>
women dressed as men	<input type="radio"/>	<input type="radio"/>
men affected than women	<input type="radio"/>	<input type="radio"/>

Figure B.2: Sample page from first pilot evaluation experiment

B.1.2 Comments

- “The first example of Bacchus and Ariadne is very hard to see and I had to guess some of the answers because I couldn’t make out the details.”
- “In the last picture, it is unlikely that they were slaves because it was painted after the end of the Civil War and the Abolition of Slavery in the US. That’s why I marked the descriptors connected to slavery as ‘bad’.”
- “Following the initial example, I’ve tried to dis-include any ungrammatical sentences.”
- “I think there is a small grammatical issue - listed words seem to be written (e.g.) work , workers rather than work, workers. I assumed this was a mistake in my answers.”
- “It’s hard to tell what counts as a good or bad description. Maybe it would be good to have a ‘don’t know’ box for options I put as bad because I didn’t know what they meant.”
- “Some phrases make sense, but because they are not grammatically correct (e.g. wrong word order or missing articles/conjunctions) I tag them as ‘bad’ description.”

B.1.3 Responses

Below are the responses of the first group of participants. The judgements are binary ('R' stands for 'Relevant' and 'N' stands for 'Non-relevant'). Group One consists of seven participants while Group Two consists of six.

Table B.1: Group One responses

“Detroit, 1943”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
speed transportation	R	R	N	N	N	N	N
automobile parked	N	R	R	N	R	R	N
buildings are white	N	N	R	N	N	N	R
time was old	R	N	N	N	R	N	N
people , buildings	N	R	R	R	R	N	N
fleet consists of cars	N	R	R	R	N	N	N
people working in city	N	R	N	R	R	N	R
buildings old	R	R	N	R	R	N	N
city north of street	N	N	N	R	N	N	R
production growth	N	N	R	N	N	N	R
buildings constructed at time	R	N	R	N	R	N	R
street carries traffic	N	R	R	N	R	R	N
“The Cotton Pickers”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
sack workers	N	N	R	N	R	R	N
women participate in work	R	N	R	R	R	N	N
gathering of work	R	N	N	N	R	N	N
workers slaves	R	R	N	N	R	N	R
woman working in field	R	R	N	N	R	N	N
field laborers	N	R	R	R	R	R	R
work , workers	R	R	R	N	N	N	N
women sold as slaves	R	R	R	R	R	R	N
women were workers	R	R	R	R	N	R	N
work done in field	R	R	N	R	N	N	N
work gathering	N	R	R	N	R	N	R
women did work	N	R	R	R	R	R	R
“Grizzly Giant Sequoia”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
tree reaching in height	N	N	R	R	R	R	N
shadow fallen	N	N	R	N	N	N	N
trees tall	N	R	N	N	R	N	N
tall with green	N	R	N	N	N	N	N
flora includes trees	N	N	N	N	N	R	N
woods is forest	N	R	N	R	N	N	R
trunk tree	R	R	N	N	R	N	N
tree grows in height	R	N	N	R	N	R	N
board made of wood	N	N	R	N	N	R	N
tree growing to tall	N	R	R	N	N	N	N
forest , california	N	R	R	R	N	N	N

moss green	N	R	N	R	R	N	R
“Loss of Schooner”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
ocean boats	R	R	N	N	R	N	R
storm brought waves	N	N	R	R	R	N	N
boats rowing	R	R	R	R	R	N	N
sinking resulted in loss	N	R	R	N	N	N	N
ships put to sea	R	R	R	R	R	R	R
painting is view	R	R	N	R	N	N	N
rescue from water	R	R	N	N	R	N	N
ocean sea	R	N	N	R	R	R	R
water or wind	R	R	N	R	R	N	N
view movement	N	N	N	R	N	N	N
storm emerged into ocean	R	R	R	R	R	N	N
sailors lost at sea	N	N	R	N	R	N	N
“Coney Island”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
summer 1934	N	R	N	R	R	N	N
beach crowded	N	R	N	R	R	N	N
american , people	N	N	R	R	R	R	R
painting figures	N	R	R	R	R	R	N
carnival figures	R	N	N	R	R	N	N
sand produced in excess	R	N	R	N	N	R	R
popular with beachgoers	N	N	R	R	N	N	N
people to beach	R	N	N	N	R	N	N
beach is on shore	R	R	N	R	R	N	N
people live on shore	N	N	R	N	N	N	N
beach covered with sand	N	R	R	R	R	R	N
beach has sand	N	R	N	R	N	N	N
“Bacchus and Ariadne”	P_0	P_1	P_2	P_3	P_4	P_5	P_6
gods men	N	R	N	N	N	N	N
women than for men	R	N	N	N	N	N	N
gods battle	N	R	N	N	N	N	N
battle , men	N	N	N	R	N	N	N
men were women	N	N	N	N	N	N	N
women than men	N	N	N	R	N	N	N
women have men	R	N	N	N	N	N	R
busy women	N	R	R	N	R	N	N
women separated from men	N	R	N	N	N	R	N
women men	N	N	N	N	N	N	N
men than for women	N	R	N	N	R	N	R
men took in battle	N	N	R	N	N	N	N

Table B.2: Group Two responses

“Detroit, 1943”	P_0	P_1	P_2	P_3	P_4	P_5
production was revival	R	R	N	N	R	N
production photography	N	N	R	N	R	N
automobiles parked	N	R	R	N	N	N
people classified as white	N	N	N	R	N	N
production begun in time	N	N	N	N	N	N
city named street	N	N	N	N	N	R
people live outside city	N	N	N	R	N	R
people old	N	R	N	N	N	N
people from outside city	N	N	R	N	R	N
speed movement	R	R	R	N	R	N
traffic in street	R	N	N	N	N	N
people spent time	R	R	R	R	N	N
“The Cotton Pickers”	P_0	P_1	P_2	P_3	P_4	P_5
women picking	R	N	N	N	R	R
gathering work	N	N	N	R	N	N
work was painting	N	R	R	R	N	R
workers employed in agriculture	N	R	R	R	R	R
work focuses on women	R	N	N	N	R	N
work published in 1876	N	R	R	R	N	R
women employed as workers	R	N	R	R	R	R
work on painting	N	R	R	N	N	R
work is landscape	R	R	N	N	N	N
work painting	N	R	R	R	N	R
women dominate field	R	R	R	N	R	R
cotton work	N	R	R	N	R	N
“Grizzly Giant Sequoia”	P_0	P_1	P_2	P_3	P_4	P_5
tall oil	R	N	R	N	N	N
forest dominated by trees	N	R	R	R	R	R
forest is nature	N	N	N	R	N	R
tree grows to height	R	R	N	R	R	R
tree growing tall	N	N	R	N	R	N
mounted at height	N	R	R	N	N	N
darkness fallen	N	N	N	N	N	R
tree found in forest	R	R	R	N	R	R
mounted painting	R	R	N	R	N	N
forest consisting of trees	R	R	R	R	R	R
painting paper	R	R	N	R	N	N
trees reach height	N	N	R	R	R	R
“Loss of Schooner”	P_0	P_1	P_2	P_3	P_4	P_5
waves damaged boats	R	R	N	N	N	N
water flows to sea	R	N	R	R	N	N

ships were at sea	R	R	R	R	N	R
sea meets ocean	R	R	N	R	N	N
sailors referred to ships	R	R	R	N	N	N
survivors of movement	N	N	R	R	R	R
view ships	R	R	R	R	N	N
rescue ships	N	N	R	N	R	N
boats rescue	R	R	R	N	R	R
ships crossing ocean	N	N	R	R	R	R
boats capsized in sea	R	R	N	N	N	R
view horizon	R	N	R	N	N	R
“Coney Island”	P_0	P_1	P_2	P_3	P_4	P_5
outdoor people	R	N	R	R	N	R
figures representing people	R	R	R	N	N	R
crowd in excess	R	N	N	N	R	N
beach located on shore	N	R	R	R	R	R
sand dunes along shore	N	N	R	N	N	R
people crowd	R	R	N	N	R	R
painting is in oil	R	N	R	N	R	N
painting tower	N	R	R	N	N	R
beach are popular	R	N	R	R	N	N
painting people	N	R	N	N	R	R
beach known for sand	R	R	R	N	N	R
summer be hot	N	R	R	R	N	R
“Bacchus and Ariadne”	P_0	P_1	P_2	P_3	P_4	P_5
women dressed as men	N	R	N	N	N	N
dionysus bacchus	N	N	N	N	N	N
men affected than women	N	R	N	N	N	N
battle against men	N	R	N	N	R	N
women than in men	R	R	N	R	N	N
battle gods	R	N	R	N	R	N
mythical gods	N	R	R	R	N	N
men were gods	R	N	N	R	N	N
women are as men	R	R	N	R	N	N
aubusson tapestry	N	R	R	N	N	R
women joined men	N	N	N	N	N	N
men keep women	R	N	R	N	R	R

B.2 Pilot experiment 2 (testing instructions)

Below are the phrases shown to the participant of the second pilot experiment, along with the responses received (in boldface).

oil painting **5**, crowd of people **10**, hot summer **4**, hot balloon **10**, american pyramid **4**, popular with people **2**, american painting **7**, american figures **5**, masses of people **5**, hot

oil 10, excess of people 7, american humor 5, american beach 8, popular beer 1, american summer 5, people crowded 7, crowded with people 7, american oil 7, american cartoon 8, hot sand 5, figures of people 2, popular stereotypes 5, popular with crowd 2, american popular 4, popular with masses 2, american beer 10, summer people 8, stereotypes of people 8, popular cartoon 5, american tower 2, popular painting 5, summer carnival 3, american stereotypes 6, crowd in excess 5, crowd figures 3, sand people 2, sand of beach 3, grotesque figures 8, humor, satire 9, american shore 6, shore of beach 10, lazy people 6, excess oil 2, american crowd 7, satire, humor 8, beach located on shore 10, oil tower 1, popular summer 6, beach with sand 10, american hot 5, people crowd 10, drunken debauchery 8, beach, marsh 10, sand painting 8, colourful figures 5, oil sand 2, summer painting 8, people in thirties 5, people painting 10, hot masses 3, beach crowded 9, cartoon figures 8, american satire 5, beach pyramid 3, outdoor painting 5, cartoon humor 3, beach tower 3, painting, oil 10, painting became popular 2, crowded with figures 8, popular confusion 5, colourful and vibrant 4, crowded beach 9, sand masses 5, colourful people 2, popular satire 6, beach people 10, popular carnival 5, figures as hitler 10, oil, beer 10, people are lazy 5, figures were popular 3, american marsh 2, pyramid tower 5, grotesque humor 7, outdoor leisure 6, popular leisure 6, hot beach 5, outdoor summer 4, flesh of swine 5, people of american 7, beer, oil 10, popular and crowded 7, beach is in summer 5, hot tower 2, cartoon carnival 7, hot people 5, beer popular 10, american excess 5, hot beer 10, summer beach 5, outdoor masses 10, people of sexuality 10, figures of painting 10, painting of tower 2, confusion over sexuality 5, crowded in summer 5, summer of 1934 2, people eat flesh 10, people visited tower 10, colourful painting 7, popular with figures 10, drunken orgy 7, tower beach 10, popular humor 6, american masses 5, popular oil 3, oil people 10, outdoor people 8, drunken people 6, confusion of people 5, figures in excess 5, people of beach 8, hot and crowded 10, marsh, sand 10, marsh and beach 10, carnival attracts people 3, painting be satire 5, oil was popular 3, summer was hot 2, people were figures 2, people located on shore 10, popular for beach 7, outdoor carnival 5, carnival balloon 10, popular american 5, popular tower 3, popular hot 3, drunken crowd 8, summer masses 10, carnival figures 5, hot crowd 3, summer sand 8, carnival crowd 7, tower figures 10, american carnival 4, people in cartoon 4, tower of people 8, satire of people 10, sexuality of people 5, hitler and people 10, tower, pyramid 10, people, masses 10, vibrant and popular 3, painting depicts figures 10, painting portrays people 10, people drank beer 3, people used oil 10, beach became popular 5, shore is popular 7, flesh tastes hot 10, people traveled in summer 3, summer are popular 10

B.3 Pilot experiment 3 (measuring inter-rater reliability)

B.3.1 Interface

Phrase Assessment Experiment

Thank you for volunteering! The experiment should take you at most 10 minutes.

You will be shown **4 pictures**. Next to each picture there will be a few phrases in note form (e.g. "house in countryside").

Imagine the following scenario: **Mary** attended a talk at an art conference and saw some art pictures in posters. She didn't know the titles of the pictures but she wanted to find them later, so she decided to write down some short phrases as memory-aids.

You will be asked to guess to what extent you believe Mary could write each one of the phrases to remember the pictures:

1. Highly Unlikely
2. Probably Not
3. Cannot Decide
4. Probably
5. Highly Likely

Along with the phrases and the picture, you will be given a list of words that came to Mary's mind when she saw the image. These words might be useful when you guess what Mary ended up doing.

An example is provided below.

Figure B.3: First page of the experiment

What Mary saw	What Mary thought	Guess what Mary wrote																																																																														
<p>Mary saw the following poster at an art conference.</p> 	<p>When Mary saw this image, the following words came to her mind:</p> <p>survivors rescue sailors ships ocean water horizon waves boats movement sea view</p>	<p>Then Mary took some notes that would help her remember the image later. Can you guess if she could write the following phrases? On a scale from 1 to 5, where:</p> <p>1 means Highly Unlikely 2 means Probably Not 3 means Cannot Decide 4 means Probably 5 means Highly Likely</p> <p>do you believe Mary could write down each phrase as an aid to her memory?</p> <p>[The underlined words are from the list on the left.]</p> <table border="1"><thead><tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr></thead><tbody><tr><td>"<u>view</u> horizon"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>sailors</u> referred to <u>ships</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>ships</u> crossing <u>ocean</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>rescue</u> <u>ships</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>sea</u> meets <u>ocean</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>boats</u> <u>rescue</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>water</u> flows to <u>sea</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>survivors</u> of <u>movement</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>boats</u> capsized in <u>sea</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>ships</u> were at <u>sea</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>view</u> <u>ships</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr><tr><td>"<u>waves</u> damaged <u>boats</u>"</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr></tbody></table>		1	2	3	4	5	" <u>view</u> horizon"	●	●	●	●	●	" <u>sailors</u> referred to <u>ships</u> "	●	●	●	●	●	" <u>ships</u> crossing <u>ocean</u> "	●	●	●	●	●	" <u>rescue</u> <u>ships</u> "	●	●	●	●	●	" <u>sea</u> meets <u>ocean</u> "	●	●	●	●	●	" <u>boats</u> <u>rescue</u> "	●	●	●	●	●	" <u>water</u> flows to <u>sea</u> "	●	●	●	●	●	" <u>survivors</u> of <u>movement</u> "	●	●	●	●	●	" <u>boats</u> capsized in <u>sea</u> "	●	●	●	●	●	" <u>ships</u> were at <u>sea</u> "	●	●	●	●	●	" <u>view</u> <u>ships</u> "	●	●	●	●	●	" <u>waves</u> damaged <u>boats</u> "	●	●	●	●	●
	1	2	3	4	5																																																																											
" <u>view</u> horizon"	●	●	●	●	●																																																																											
" <u>sailors</u> referred to <u>ships</u> "	●	●	●	●	●																																																																											
" <u>ships</u> crossing <u>ocean</u> "	●	●	●	●	●																																																																											
" <u>rescue</u> <u>ships</u> "	●	●	●	●	●																																																																											
" <u>sea</u> meets <u>ocean</u> "	●	●	●	●	●																																																																											
" <u>boats</u> <u>rescue</u> "	●	●	●	●	●																																																																											
" <u>water</u> flows to <u>sea</u> "	●	●	●	●	●																																																																											
" <u>survivors</u> of <u>movement</u> "	●	●	●	●	●																																																																											
" <u>boats</u> capsized in <u>sea</u> "	●	●	●	●	●																																																																											
" <u>ships</u> were at <u>sea</u> "	●	●	●	●	●																																																																											
" <u>view</u> <u>ships</u> "	●	●	●	●	●																																																																											
" <u>waves</u> damaged <u>boats</u> "	●	●	●	●	●																																																																											

Figure B.4: One of the four image pages

B.3.2 Responses

‘P0’ stands for ‘Participant 0’, ‘P1’ for ‘Participant 1’ and ‘Me’ for myself.

Table B.3: Pilot experiment

“The Cotton Pickers”	P_0	P_1	Me
women picking	5	5	4
gathering work	4	4	2
work was painting	2	1	1
workers employed in agriculture	4	4	5
work focuses on women	3	1	2
work published in 1876	2	5	4
women employed as workers	5	4	5
work on painting	2	3	2
work is landscape	1	1	2
work painting	3	2	1
women dominate field	2	1	4
cotton work	5	4	1
“Coney Island”	P_0	P_1	Me
outdoor people	2	5	4
figures representing people	3	4	1
crowd in excess	4	4	2
beach located on shore	2	1	4
sand dunes along shore	2	1	4
summer be hot	1	1	1
painting is in oil	2	4	4
beach known for sand	1	2	2
beach are popular	2	4	1
painting people	3	3	2
people crowd	2	4	2
painting tower	1	1	1
“Grizzly Giant Sequoia”	P_0	P_1	Me
tall oil	2	1	1
forest dominated by trees	3	3	4
forest is nature	1	1	2
tree grows to height	2	4	2
tree growing tall	5	5	4
mounted at height	1	2	1
darkness fallen	2	2	4
tree found in forest	5	4	4
mounted painting	1	2	4
forest consisting of trees	3	4	5
painting paper	1	1	2
trees reach height	1	4	2

“Loss of Schooner”	P_0	P_1	Me
waves damaged boats	3	4	5
water flows to sea	1	1	2
ships were at sea	3	5	4
sea meets ocean	1	1	2
sailors referred to ships	3	2	1
rescue ships	5	5	5
view ships	3	1	1
survivors of movement	1	1	2
boats rescue	3	4	5
ships crossing ocean	4	3	4
boats capsized in sea	5	4	2
view horizon	3	2	1

B.4 Main experiment

B.4.1 Screenshot

Image 1 of 4

What Mary *saw*

Mary saw the following poster at an art conference.

"Proserpine" (1874)
by Dante Gabriel Rossetti
Tate Modern, London



Guess what Mary *wrote*

Then Mary took some notes that would help her remember the image later. Can you guess if she could write the following phrases? On a scale from 1 to 5, where:

1 means **Highly Unlikely**
2 means **Probably Not**
3 means **Cannot Decide**
4 means **Probably**
5 means **Highly Likely**

do you believe Mary could write down each phrase as an aid to her memory?

	1	2	3	4	5
"beautiful woman"	●	●	●	●	●
"painting is portrait"	●	●	●	●	●
"painting of woman"	●	●	●	●	●
"rossetti painting"	●	●	●	●	●
"blue dress"	●	●	●	●	●
"portrait gallery in london"	●	●	●	●	●
"dante gabriel rossetti"	●	●	●	●	●
"beautiful women"	●	●	●	●	●
"portrait of wife"	●	●	●	●	●
"woman dressed in dress"	●	●	●	●	●
"portrait of woman"	●	●	●	●	●
"british library in london"	●	●	●	●	●
"allegorical painting"	●	●	●	●	●
"hands of british"	●	●	●	●	●
"british museum in london"	●	●	●	●	●
"beautiful young woman"	●	●	●	●	●
"woman in dress"	●	●	●	●	●
"blue velvet"	●	●	●	●	●
"dress worn by women"	●	●	●	●	●
"elegant dress"	●	●	●	●	●
"dramatic art in london"	●	●	●	●	●
"london with his wife"	●	●	●	●	●
"woman identified as wife"	●	●	●	●	●
"british women"	●	●	●	●	●
"wife of british"	●	●	●	●	●
"portrait painting"	●	●	●	●	●
"portrait engraved from painting"	●	●	●	●	●

Can you suggest **your own** phrases that describe aspects of this picture?
Please write at least three phrases (2-4 words per phrase).

Figure B.5: One of the four image pages displayed in the main evaluation experiment

B.4.2 Scores per image per system

Below are the scores that each phrase of both systems received in each image, averaged over all nine participants.

Table B.4: Responses of main experiment

“Loss of Schooner” (Ngram)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
water to prevent ships	1.0	1	1	1	1	1	1	1	1	1
ocean and mediterranean sea	1.56	1	1	2	4	1	1	2	1	1
ocean and caribbean sea	1.78	1	1	2	4	1	1	4	1	1
sea and indian ocean	1.89	1	1	2	4	1	1	4	2	1
sea and atlantic ocean	1.67	1	1	2	4	1	1	3	1	1
ships at sea	3.67	5	2	4	5	2	2	5	4	4
loss of water	1.33	1	1	1	1	2	2	2	1	1
wind and water	2.89	4	1	3	4	1	3	4	2	4
movement of water	1.78	2	1	2	4	1	1	2	2	1
ships and boats	3.44	5	2	3	5	3	2	4	4	3
sea water	2.0	2	1	2	5	1	2	2	1	2
storm water	1.78	1	2	3	1	1	1	4	2	1
ocean view	1.78	2	1	3	2	2	1	3	1	1
ocean water	2.11	1	1	2	5	1	2	3	3	1
water loss	1.22	1	1	1	1	2	1	2	1	1
“Loss of Schooner” (POS-Dep)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
ships caught in storm	4.33	4	5	5	4	4	4	4	4	5
ships lost at sea	3.67	5	2	5	5	5	2	4	1	4
sailors lost at sea	3.22	5	2	5	4	3	2	4	1	3
ships lost in storm	3.56	4	4	5	1	5	2	5	3	3
ships damaged by storm	3.67	4	4	4	4	5	2	4	4	2
wind and waves	3.33	4	2	3	4	2	3	5	4	3
movement of water	1.78	2	1	2	4	1	1	2	2	1
loss of ships	2.67	1	1	3	5	5	1	4	1	3
ships at sea	3.67	5	2	4	5	2	2	5	4	4
boats and ships	3.0	4	2	2	4	3	2	4	4	2
ocean waves	2.78	5	2	1	2	1	3	5	4	2
storm waves	2.89	5	2	2	4	2	3	3	4	1
rescue ships	3.89	4	2	3	4	5	4	5	4	4
sea waves	3.44	5	2	2	4	1	4	5	4	4
ocean storm	3.56	5	4	1	4	2	4	4	4	4
“Proserpine” (Ngram)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
british museum in london	2.11	4	1	1	5	1	1	1	4	1
portrait gallery in london	2.89	4	1	2	5	5	2	3	3	1
london with his wife	1.33	1	1	1	4	1	1	1	1	1

dramatic art in london	2.67	4	1	4	5	4	1	2	1	2
british library in london	1.33	4	1	1	1	1	1	1	1	1
beautiful young woman	3.22	4	2	1	5	3	3	4	3	4
dante gabriel rossetti	3.78	5	1	5	5	5	2	5	5	1
hands of british	1.33	1	1	1	4	1	1	1	1	1
wife of british	1.33	1	1	1	4	1	1	1	1	1
portrait of woman	4.22	5	4	4	5	5	4	4	3	4
beautiful woman	3.22	5	2	1	5	3	3	4	3	3
beautiful women	2.11	1	1	1	1	3	3	4	4	1
british women	1.89	1	1	1	5	3	1	2	2	1
blue velvet	3.67	4	1	3	5	5	3	4	4	4
portrait painting	2.56	2	2	1	4	3	2	5	3	1

“Proserpine”
(POS-Dep)

	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
woman dressed in dress	2.0	4	1	3	1	2	2	1	1	3
woman identified as wife	1.67	1	1	2	4	1	2	2	1	1
dress worn by women	2.22	2	1	1	4	3	4	1	2	2
portrait engraved from painting	1.22	1	1	1	3	1	1	1	1	1
portrait of woman	4.22	5	4	4	5	5	4	4	3	4
woman in dress	3.22	5	1	2	4	3	3	3	4	4
portrait of wife	2.33	3	2	3	4	1	3	1	1	3
painting of woman	4.11	5	2	4	5	5	4	5	4	3
painting is portrait	2.11	2	1	2	1	3	3	1	3	3
blue dress	3.89	5	1	3	5	5	3	5	4	4
beautiful woman	3.22	5	2	1	5	3	3	4	3	3
rossetti painting	4.22	5	4	5	5	5	2	5	5	2
elegant dress	3.22	4	1	2	5	5	3	2	3	4
allegorical painting	2.89	4	1	3	3	3	1	4	4	3

“Sunlight”
(Ngram)

	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
american society of landscape	1.11	1	1	1	1	1	2	1	1	1
water level in lake	1.11	1	1	1	1	1	2	1	1	1
water level of lake	1.11	1	1	1	1	1	2	1	1	1
cliff richard and shadows	1.44	1	1	1	4	1	2	1	1	1
lake of two mountains	1.44	1	1	2	1	2	1	1	1	3
lady of lake	3.11	4	2	4	4	3	3	5	1	2
water from lake	1.67	2	1	2	1	3	1	2	1	2
lady in waiting	3.33	4	4	2	5	4	2	3	2	4
woman in white	4.44	5	2	5	5	5	4	5	5	4
wind and water	1.89	2	1	2	2	1	1	3	4	1
american woman	1.89	4	1	2	2	3	1	2	1	1
white light	2.56	4	1	4	5	2	1	2	3	1
white mountains	1.44	3	1	1	3	1	1	1	1	1
bright light	2.44	4	1	3	5	2	1	1	4	1
white woman	3.44	5	2	4	4	3	2	5	4	2

“Sunlight” (POS-Dep)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
lake is in summer	1.56	1	1	2	1	3	1	1	3	1
woman dressed in dress	1.67	2	1	3	1	3	1	1	1	2
lake surrounded by mountains	2.67	2	1	4	5	2	2	2	2	4
lake filled with water	1.78	2	1	1	2	2	2	1	4	1
water is in summer	1.22	1	1	2	1	1	1	1	2	1
woman in dress	3.56	4	2	4	4	5	2	4	5	2
light in sky	1.78	1	2	1	5	1	1	1	3	1
water and sky	2.67	5	1	3	4	3	2	1	4	1
profile of woman	3.33	4	2	3	4	5	2	5	4	1
bright and sunny	2.56	4	1	4	4	2	2	1	4	1
bright light	2.44	4	1	3	5	2	1	1	4	1
white dress	3.78	4	2	4	5	5	2	3	5	4
bright sunlight	3.0	4	2	4	5	5	1	1	4	1
white cloud	1.67	1	1	3	1	1	1	1	4	2
white clouds	3.0	5	1	3	5	4	1	1	5	2

“House in Provence” (Ngram)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
village in howmeh rural	1.11	1	1	1	1	1	1	1	2	1
country at 1996 summer	1.0	1	1	1	1	1	1	1	1	1
country at 2000 summer	1.0	1	1	1	1	1	1	1	1	1
country at 2004 summer	1.0	1	1	1	1	1	1	1	1	1
village in golestan rural	1.33	1	2	1	3	1	1	1	1	1
english and french	1.33	5	1	1	3	1	1	1	1	1
french and english	1.78	4	1	2	4	1	1	1	1	1
summer and autumn	1.0	1	1	1	1	1	1	1	1	1
blue ridge mountains	2.33	3	1	1	4	2	3	2	4	1
english and scottish	1.56	4	1	2	1	1	1	1	2	1
country house	4.0	5	4	5	4	4	2	4	4	4
blue mountains	3.56	5	4	4	4	3	3	2	4	3
blue sky	2.78	4	1	2	4	2	2	5	4	1
irish house	1.56	4	2	1	1	2	1	1	1	1
village green	2.11	3	1	3	4	4	1	1	1	1

“House in Provence” (POS-Dep)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
landscape consists of mountains	2.11	1	1	1	2	3	2	1	4	4
house surrounded by trees	3.44	3	2	4	4	3	4	3	4	4
landscape consists of hills	2.0	1	1	1	4	3	3	1	2	2
house situated in village	1.44	1	1	1	2	3	1	1	2	1
village surrounded by trees	1.22	2	1	1	1	1	1	1	1	2
landscape of trees	1.89	1	1	2	2	2	2	1	4	2
landscape of hills	2.56	2	4	4	2	2	3	1	2	3
house in provence	3.67	5	2	5	4	5	2	4	4	2

orchard of trees	1.78	1	1	4	2	1	2	2	2	1
fields and trees	2.33	4	1	3	4	3	1	1	3	1
rural landscape	3.44	5	4	3	2	3	2	5	4	3
green trees	2.44	2	1	4	4	2	2	2	4	1
village green	2.11	3	1	3	4	4	1	1	1	1
green mountains	1.11	1	1	1	1	2	1	1	1	1
english countryside	1.89	5	1	1	1	2	1	1	4	1
“House in Provence” (Parallel)	Mean score	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
white house	3.0	5	1	3	1	3	3	4	4	3
blue mountains	3.56	5	4	4	4	3	3	2	4	3
simple house	2.78	1	2	3	5	3	2	4	4	1
green trees	2.44	2	1	4	4	2	2	2	4	1
grey mountains	2.33	4	2	2	1	3	2	2	4	1
house surrounded by trees	3.44	3	2	4	4	3	4	3	4	4
painting of house	3.78	5	2	5	4	4	3	5	4	2
house in countryside	4.0	4	5	4	4	5	2	4	4	4
painting of landscape	3.11	4	2	3	4	2	2	5	4	2
landscape with mountains	3.78	4	4	5	2	3	3	5	5	3
countryside with mountains	3.11	4	2	3	2	3	3	4	4	3

B.4.3 Phrases suggested by participants

Below are the raw (left column) and normalised (right column) phrases submitted by participants. Phrases whose *both* end words are tags in the image’s tag cloud have these two words emphasised. If at most one end-word is a tag, the normalised phrases are presented below without any formatting.

Table B.5: User-submitted phrases and their conversion

“House in provence”

“Landscape and house”	→	– <i>landscape and house</i>
“House in countryside”	→	– <i>house in countryside</i>
“French countryhouse”	→	– <i>french countryhouse</i>
“Dead trees”	→	– <i>dead trees</i>
“green field”	→	– <i>green field</i>
“house with few windows”	→	– <i>house with windows</i> – <i>few windows</i>
“French country house”	→	– <i>french house</i> – <i>country house</i>
“blue mountains”	→	– <i>blue mountains</i>
“simple landscape”	→	– <i>simple landscape</i>
“house in blue mountains”	→	– <i>house in mountains</i> – <i>blue mountains</i>
“painting country house provence”	→	– <i>country house</i>
“Cézanne painting landscape”	→	– <i>cezanne painting</i> – <i>painting landscape</i>
“Simple house in countryside”	→	– <i>simple house</i> – <i>house in countryside</i>
“Blue hills and sky”	→	– <i>blue hills</i> – <i>blue sky</i> – <i>hills and sky</i>
“Winter and autumn”	→	– <i>winter and autumn</i>
“Varied terrains”	→	– <i>varied terrains</i>
“hidden house”	→	– <i>hidden house</i>
“rural scene ”	→	– <i>rural scene</i>
“house by mountains”	→	– <i>house by mountains</i>
“white house in valley in front of mountains”	→	– <i>white house</i> – <i>house in valley</i> – <i>valley in front of mountains</i> (> 4 words)
“house in front of distant mountains”	→	– <i>house in front of mountains</i> (> 4 words) – <i>distant mountains</i>
“mountains behind trees surrounding a white country house”	→	– <i>mountains behind trees</i> – <i>trees surrounding house</i> – <i>white house</i> – <i>country house</i>
“Landscape house mountains”	→	–
“house blue mountains landscape”	→	–

“House trees mountains
distance” → -

“Loss of Schooner”

“Ship lost in storm” → - ship lost in storm

“Rescue boat” → - rescue boat

“sunk ship and survivors” → - sunk ship
- ship and survivors

“Ship sinking” → - ship sinking

“rough sea” → -

“high waves” → - high waves

“stormy, grey” → - stormy, grey

“thunder storm” → - thunder storm

“sinking ships” → - *sinking ships*

“boats distress at sea” → - *boats* distress at *sea*

“shipwreck at sea” → - *shipwreck* at *sea*

“stormy rescue

mission ocean” → - stormy rescue
- rescue mission

“Turbulent waves

of the sea” → - turbulent waves
- *waves* of *sea*

“Stormy sky ” → - stormy sky

“Deep dark water” → - deep water
- dark water

“Damaged and
sinking boats” → - damaged boats
- *sinking boats*

“sinking ship” → - sinking ship

“ships in storm” → - *ships* in *storm*

“abandon ship” → - abandon ship

“ships in stormy sea” → - *ships* in *sea*
- *stormy sea*

“rescue from sinking
due to storm” → - *rescue* from *sinking*
- *sinking* due to *storm*

“ship sunk by
large waves” → - ship sunk
- sunk by waves
- large waves

“Ship rescue” → - ship rescue

“Sinking ship” → - sinking ship

“Storm ship sinking” → - ship sinking

“Sunlight”

“Woman in lake
landscape” → - *woman* in *landscape*
- *lake landscape*

“White dress woman” → - *white dress*

“Woman looking into
the horizon” → - woman looking into horizon

“woman looking against sunlight”	→	– <i>woman</i> looking against <i>sunlight</i>
“dress flowing in air”	→	– dress flowing in air
“elegant lady”	→	– elegant lady
“grassy hill overlooking lake”	→	– grassy hill – <i>hill</i> overlooking <i>lake</i>
countryside	→	–
“muted colours, oil painting”	→	– muted colours – oil painting
“lady white dress lake”	→	– <i>white dress</i>
“woman in summer sun”	→	– woman in sun – summer sun
“woman in white lake”	→	– <i>woman</i> in <i>lake</i> – <i>white lake</i>
“Woman looking beyond a lake”	→	– <i>woman</i> looking beyond <i>lake</i>
“White lace dress with a high neck”	→	– <i>white dress</i> – lace dress – high neck – dress with neck
“Bright and sunny”	→	– <i>bright</i> and <i>sunny</i>
“Neatly tied hair”	→	– tied hair
“woman on hill”	→	– <i>woman</i> on <i>hill</i>
“woman looking at water”	→	– <i>woman</i> looking at <i>water</i>
“lady in white on hill”	→	– <i>lady</i> in <i>white</i> – <i>lady</i> on <i>hill</i>
“woman in white dress at waterfront”	→	– <i>woman</i> in <i>dress</i> – <i>white dress</i> – woman at waterfront
“woman on a hillside by the waterfront”	→	– <i>woman</i> on <i>hillside</i> – woman by waterfront
“woman in white dress staring down the lake”	→	– <i>woman</i> in <i>dress</i> – <i>white dress</i> – <i>woman</i> staring down <i>lake</i>
“Woman looks into distance”	→	– woman looks into distance
“White dress woman lake”	→	– <i>white dress</i>
“Profile woman lake grass”	→	–

“Proserpine”

“Blue dress woman”	→	– <i>blue dress</i>
“Rossetti’s blue woman”	→	– <i>rossetti</i> woman – <i>blue woman</i>

“Melancholic woman”	→	– melancholic woman
“black-haired woman”	→	– black-haired woman
“holding pomegranate”	→	– holding pomegranate
“mysterious facial expression”	→	– mysterious expression – facial expression
“long blue dress”	→	– long dress – <i>blue dress</i>
“painting of woman”	→	– <i>painting of woman</i>
“woman in blue dress”	→	– <i>woman in dress</i> – <i>blue dress</i>
“rosetti fruit in hand”	→	– fruit in hand
“woman holding hand”	→	– woman holding hand
“beauty in blue dress”	→	– beauty in dress – <i>blue dress</i>
“Lost in thought”	→	– lost in thought
“Rich velvet silhouette”	→	– rich velvet
“Thick black hair”	→	– thick hair
“Strong dark colours”	→	– strong colours – dark colours
“woman and fruit”	→	– <i>woman and fruit</i>
“painting of woman”	→	– <i>painting of woman</i>
“painting eve metaphor”	→	– painting eve – eve metaphor
“sad woman in dark dress eating a fruit”	→	– sad woman – <i>woman in dress</i> – dark dress – <i>woman eating fruit</i>
“sad woman in velvet dress”	→	– sad woman – <i>woman in dress</i> – <i>velvet dress</i>
“sad woman eating a fruit”	→	– sad woman – <i>woman eating fruit</i>
“Woman Blue Velvet Pomegranate”	→	– <i>blue velvet</i>
“Portrait Woman Holding Pomegranate”	→	– <i>woman holding pomegranate</i>
“Beautiful woman blue dress”	→	– <i>beautiful woman</i> – <i>blue dress</i>