Extracting Perceived Landscape Properties from Text Sources

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat)

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Olga Koblet

aus

Russland

Promotionskommission

Prof. Dr. Ross Stuart Purves (Vorsitz)

Prof. Dr. Christian Berndt

Prof. Dr. Peter Verburg

Prof. Dr. Devis Tuia

Zürich 2020

Faculty of Science
University of Zurich

Extracting Perceived Landscape Properties from Text Sources

Olga Koblet
Geocomputation Unit
Department of Geography
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

"The joys of being able to seek out the less spectacular and easier hills have been evident in a few walks recently. I have followed broad grassy ridges, often alone for most of the day, save Hyperdog Moss, no other walkers in sight. Relatively easy walking, lovely views, that invigorating feeling of the wind on your face, and the silence. Not the silence as defined by the dictionary, but the silence of the hills - the wind, the occasional call of a bird or a sheep, the sound of my breathing as I labour uphill. For me this is the silence of peace, the silence that allows you to empty your mind of thoughts and by default allow the senses to roll in and out of your mind. Sight, sound, smell, taste and touch. The views, the birds, the bracken, the water from a stream, the contrast between rock and grass under your trail shoe or boot."

Travel diary entry of FellBound: https://fellbound.blogspot.com/2014/11/

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to Prof. Dr. Ross S. Purves for supervising my thesis. His kindness, professional competence, readiness and availability for any type of discussion, encouragement and incredible patience made this work possible. For various specific pieces of information, interesting suggestions, as well as general support during my thesis I would like to thank my PhD committee members Prof. Dr. Christian Berndt, Prof. Dr. Peter Verburg and Prof. Dr. Devis Tuia. I would also like to express my special thanks to the Graduate School in Geography for the support, organisation of useful courses and social events.

I am grateful to the members of the Geocomputation group for their feedback on numerous dry runs, for teaching me about reproducible research and scientific ethics and for broadening my cultural and scientific horizons. I am especially grateful to Maximilian for his valuable feedback on the thesis. I would also like to express my special thanks to Ben, Jo, Flurina, Curdin, Michele, Diego and Oliver for helping me not only in my research, but also in my general understanding of how academia works and for opening up new opportunities.

I would like to thank Michelle, Raha, Julia, Annina, Sascha and James for taking care of the fun part during these four years and for controlling that I move gradually from one PhD phase to another without stacking too long in the frustration and apathy ones. I would like to extend thanks to the rest of the former and current members of the GIScience unit for making this journey interesting and inspiring.

I am grateful to my family and friends for their love, support, encouragement and patience, especially to Anna, Natalya and Iris. Above all, I would like to thank Thomas, it would not have been possible to accomplish this project without you!

Finally, I would like to thank Alexander, Sergey, Yurate, Elena Aleksandrovna, Victor, Esra and Ivan for their tremendous support during other phases of my education, which made it possible for me to walk towards J-82 on 15th of September 2015.

Zurich, September 2019

SUMMARY

In parallel with the emergence of new data sources and the re-discovery of existing sources, such as written first-person narratives available in travel reports and diaries, is an increasing realisation of the importance of capturing bottom-up ways of experiencing landscapes. This recognition is reflected in different policy works including overarching frameworks European Landscape Convention and Millennium Ecosystem Assessment, and local ones, such as Landscape Character Assessment in England and Scotland (LCA) and the Swiss Landscape Monitoring Program. Important challenges for these frameworks are how to include multiple perspectives of landscape perception and how to integrate different senses including sound and smell experiences, memories and associations, and experiential perceptions such as touch and feel.

The proliferation of new data in the form of natural language has brought with it a need for robust and reproducible workflows allowing extraction and classification of descriptions referring to perceived landscape properties. Therefore, the overall aim of this thesis is to explore the potential of written first-person narratives for landscape assessment and to develop methodological workflows, which can extract and classify information containing visual, aural and olfactory perception as well as tranquillity from natural language.

To approach this aim, we set out a series of experiments in Great Britain and the English Lake District, first, demonstrating to what degree landscape scenicness can be modelled purely as a function of language (Publication 1), second, extracting and classifying information of other senses from written first-person narratives (Publications 2, 3, 4) and exploring temporal changes in landscapes, in perception and in their polarity (Publication 3). Lastly, we created a spatial corpus of written first-person narratives and assessed if the collected information is useful for practice (Publication 4).

Our model based on written first-person narratives was able to explain 52% of the variation in scenicness, comparable to models using more traditional approaches of interviews and participatory methods, land cover data and social media, demonstrating that textual descriptions are feasible to use in studies of landscape perception. From these descriptions we were able to extract more than 8000 explicit references to aural perception in Great Britain, which accounts for a small percentage of all descriptions (ca. 0.25%), but in its absolute value exceeds what can be collected using interviews and does not intrude into the experiences of people. Estimation of polarity gave an additional level of understanding of descriptions classified into different types of sound emitter with no clear distinction of, for example, anthrophony being more negative than biophony or

geophony, contrary to the statements in literature that natural sounds have positive connotations. The majority of extracted descriptions in Great Britain (ca. 59%) referred to the perceived absence of sound, therefore, we undertook a detailed study of intertwined visual and aural perception reflected in the concept of tranquillity concentrating on the Lake District region. To do so we used historical and contemporary corpora and a combination of micro- and micro-analysis, allowing us, first, to develop a taxonomy of tranquillity as encoded in natural language and, second, to explore the changes in descriptions of the Lake District, such as an overall decline of mentions of total silence and increase in the references to tranquillity as a contrast to anthropogenic intrusions. By mapping our results we were able to demonstrate that spatial modelling based on proximity to potential noise emitters as a proxy of tranquillity disturbance does not reveal tranquillity pockets close to transportation arteries, which emerged through our analysis.

The aforementioned results were obtained from textual resources partly unique to Great Britain (ScenicOrNot and Geograph datasets). To demonstrate the transferability of our results to other territories, we created a workflow allowing collection of first-person narratives for a region of interest. Using this workflow we were able to collect almost 7000 rich first-person narratives of the Lake District, comprising ca. 8 million of words. From these descriptions we extracted, classified and linked to space more than 28000 references to visual perception, almost 1500 to aural perception and tranquillity and 78 explicit references to olfactory experiences. We explored these descriptions using four levels of granularity: Great Britain, the Lake District, areas of distinctive character as used in LCA, and individual named landscape elements. We presented the resulting dataset to the Lake District National Park Authority and an important local pressure group Friends of the Lake District, who gave their feedback on strengths and weaknesses of our approach and explicitly confirmed its value for LCA and other monitoring activities of the National Park.

Overall, our results demonstrated that written first-person narratives are a valuable source of landscape perception complementary to field-based studies, since they contain information about different types of perceived landscape properties that can be extracted and classified and extend temporal coverage as demonstrated through analysis of the historical corpus. However, the possibility to 'go back in time' should not necessarily mean several centuries; it can be useful also for shorter time periods, when, for example, no interviews were conducted for a certain area.

Important limitations of our approach from the data source point of view include a potential bias towards people who enjoy writing and a potential over/under-representation of certain groups based on other criteria. From the methodological point of view, despite the advances of natural languages processing tools and techniques, natural language remains a challenging source of information for analysis, including problems related to disambiguation of

words that can be used in different senses, detection of metaphorical and ironical phrases and differentiation between mentioned locations that are visited or simply seen. Further research has to be done, first, to improve these methods and, second, to develop clear criteria that allow the assessment of how balanced a corpus of written first-person narratives is.

We see great potential of our results for several fields of studies including GIScience, landscape studies, digital humanities and tourism studies. Information available in textual sources can be scaled up to cover large spatial extents, offering GIScience, first, to add an additional dimension of human experiences in the research related to, for example, sense of place and delineation of cognitive regions, and, second, to increase the participation in the production of spatial information and knowledge. Methods used in this work can be extended to explore other landscape-related concepts, which are likely to be captured in the natural language, including concepts of wilderness and naturalness. Thus, we see value in integrating our results into a more general landscape preference model based on written first-person narratives and textual analysis. The corpus of first-person perception in the Lake District created in this work contains a plethora of writers and viewpoints, and allows researchers in the digital humanities to continue exploring the words (e.g., 'sublime', 'picturesque') and concepts they refer to (e.g., 'scenery', 'manner') as selected by contemporary authors to describe their affections towards landscapes. This information, as shown throughout our work, gives an additional level of understanding of the ways writing of the forebears has influenced our landscape perception today, and suggests to explore deeper to which extent modern day tourism follows the foundations laid by Dorothy and William Wordsworth in the Romantic era and by Alfred Wainwright throughout the 20th century.

This thesis is presented in two parts: a synthesis, describing the project as a whole, and the following 4 publications, which are found in the appendix.

Publication 1:

Chesnokova, O., Nowak, M., and Purves, R.S., 2017. A crowdsourced model of landscape preference. In: E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore, eds. 13th International Conference on Spatial Information Theory (COSIT 2017). Leibniz International Proceedings in Informatics, 19:1–19:13.

Publication 2:

Chesnokova, O. and Purves, R.S., 2018. From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text. Applied Geography, 93, 103–111.

Publication 3:

Chesnokova, O., Taylor, J.E., Gregory, I.N., and Purves, R.S., 2019. Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District. International Journal of Geographical Information Science, 33:12, 2430-2454.

Publication 4:

Koblet, O. and Purves, R.S., 2020. From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally. Landscape and Urban Planning, Volume 197, 103757.

CONTENTS

Acknowledgments v
Abstract vii
List of Figures xiii
List of Tables xiv

1 INTRODUCTION 1
1.1 Motivation and relevance 1
1.2 Thesis overview 3
2 BACKGROUND 5
2.1 Key concepts of landscape perception 5
2.1.1 Landscape and language 5
2.1.2 Visual perception and nature conservation in Great Britain 6
2.1.3 Landscape Character Assessment framework 8
2.1.4 Soundscapes 11
2.1.5 Tranquillity 12
2.1.6 Olfactory perception 13
2.2 Overview of methods used to capture perceived landscape properties 14
2.3 Methods of text analysis 17
2.3.1 Creating text corpora 19
2.3.2 Pre-processing 20
2.3.3 Ways of dealing with unstructured text 20
2.3.4 Extracting relevant text snippets 22
2.3.5 Classification 23
2.3.6 Assigning texts to space 25
2.3.7 Evaluation measures 29
2.3.8 User-generated biases 29
2.4 Research gap and research questions 30
3 FROM TEXT SOURCES TO PERCEIVED LANDSCAPE PROPERTIES 33
3.1 Study and case study areas 33
3.2 Datasets used in this work 36
3.3 Creating a text corpus 38
3.4 Language as a scenicness predictor 43
3.5 Perceived landscape properties 44
3.5.1 Visual perception 46
3.5.2 Aural perception 47
3.5.3 Tranquillity 49
3.5.4 Olfactory perception 51

3.5.5 Assigning texts to space 51
3.6 Landscape characterisation 53
3.6.1 Great Britain 54
3.6.2 Lake District 57
3.6.3 Areas of distinctive character 61
3.6.4 Individual landscape elements 65
3.6.5 Face validity 66
4 DISCUSSION 71
4.1 Textual corpus for first-person landscape perception (RQ1) 71
4.2 Variation of perceived landscape properties (RQ2, RQ3) 74
4.3 Characterisation and comparison of landscapes (RQ4) 77
4.4 Towards landscape monitoring and landscape assessment (RQ5) 79
4.5 Overall limitations 81
4.6 New possibilities and ways forward 82
5 CONCLUSIONS AND OUTLOOK 85

REFERENCES 87

A APPENDIX: PUBLICATION 1 107
B APPENDIX: PUBLICATION 2 121
C APPENDIX: PUBLICATION 3 131
D APPENDIX: PUBLICATION 4 157

LIST OF PRESENTATIONS, PUBLICATIONS AND WORKSHOPS 175

LIST OF FIGURES

Figure 1.1 Schematic overview of the interrelation between research objectives and publications. 4
Figure 2.1 Landscape as defined in the Landscape Character Assessment [*Tudor*, 2014]. 9
Figure 2.2 Lake District National Park LCA report [*Watkins*, 2008]. 10
Figure 2.3 Tranquillity map [*Hewlett et al.*, 2017]. 13
Figure 2.4 Maps of smellscapes. 14
Figure 2.5 Example of dependency parser results. 21
Figure 2.6 Example of word sense disambiguation. 22
Figure 2.7 Example of referent ambiguity. 25
Figure 2.8 Evaluation measures. 29
Figure 2.9 Term profiles for Flickr tags. 30
Figure 3.1 Our study area: The English Lake District. 34
Figure 3.2 Fragment of the *Pictorial Guide to the Lakeland Fells* [*Wainwright*, 1966]. 35
Figure 3.3 Typical landscapes of the Lake District National Park. 36
Figure 3.4 Corpus creation workflow. 41
Figure 3.5 Maps of the scenicness prediction results. 45
Figure 3.6 Overall workflow used to extract and classify descriptions of landscape perception. 45
Figure 3.7 Counts of photographs rated in the ScenicOrNot project. 46
Figure 3.8 Overall workflow for visual perception. 47
Figure 3.9 Overall workflow for aural perception. 48
Figure 3.10 Overall workflow for perceived tranquillity. 50
Figure 3.11 Final workflow for aural perception and tranquillity. 51
Figure 3.12 Overall workflow for olfactory perception. 52
Figure 3.13 Assigning texts to space. 53
Figure 3.14 Average scenicness for 150 most frequent nouns. 54
Figure 3.15 Aggregated number of descriptions related to absence of sound (macro-analysis) with selected descriptions for micro-analysis. 55
Figure 3.16 Proportion of descriptions according to sentiment values. 56
Figure 3.17 150 most frequent tokens describing geophony and biophony in Great Britain. 57
Figure 3.18 Predicted scenicness of the Lake District. 58

Figure 3.19 Number of sentences per tranquillity class in the Lake District. 60
Figure 3.20 150 most frequent nouns of contrasting sounds in the Lake District. 60
Figure 3.21 Comparison of the map of relative tranquillity [*MacFarlane et al.*, 2004] and types of tranquillity extracted from Geograph Lake District. 61
Figure 3.22 Spatial distribution of perceived sounds. 62
Figure 3.23 Interface of the maps presented to the expert group. 67

LIST OF TABLES

Table 2.1 Ways of handling unstructured text. 21
Table 2.2 Exemplary works of geographic applications. 27
Table 2.2 Exemplary works of geographic applications. 28
Table 3.1 Examples of Geograph contributions with the ScenicOrNot votes. 37
Table 3.2 Search terms and their types. 40
Table 3.3 Corpus of first-person perception in the Lake District. 42
Table 3.4 Textual corpora used in the project. 42
Table 3.5 External data sources used in the project. 43
Table 3.6 Ten most frequent syntactic pairs from scenic and unattractive lexicons. 58
Table 3.7 Summary of extracted descriptions of aural perception and tranquillity (Lake District). 59
Table 3.8 Summary of extracted descriptions of aural perception and tranquillity (individual areas). 64
Table 3.9 Ten most frequent mentions of landscape elements. 66
Table 3.10 Summary of the SWOT evaluation. 69
Table 4.1 Categories of different groups of landscape users. 73

1

INTRODUCTION

1.1 MOTIVATION AND RELEVANCE

Landscapes, and the way we value them, are not constants, but an ever-changing combination of interactions between people, culture and environment. Debates over 'beautiful', 'picturesque' and 'sublime' landscapes have been continuous since the beginning of the Romantic era towards the end of the 18th century, which led to the recognition of the most beautiful landscapes worth protecting [*Selman and Swanwick*, 2010]. Since then, three important developments have taken place. First, perceived landscape properties are officially included in the landscape monitoring of many countries, including 41 European countries that have signed the European Landscape Convention (ELC) [*Council of Europe*, 2000]. Second, there has been a shift from the monitoring and protection of merely the 'best' landscapes, to the inclusion of all landscapes without exceptions [*Swanwick and Fairclough*, 2018]. Lastly, landscapes are not limited any longer to their visual aspects, but considered to be a combination of different senses [*Tudor*, 2014].

In 2005, a global assessment program – Millennium Ecosystem Assessment (MEA) – was launched. Its main purpose is to monitor changes in ecosystems and assess their impact on human well-being. MEA combines understanding of landscape as a physical feature and as a cultural construct. In addition to provisioning, regulating and supporting services, which mostly rely on physical properties of landscapes, MEA identifies Cultural Ecosystem Services, 'tightly bound to human values' and states that 'perceptions of cultural services are more likely to differ among individuals and communities than, say, perceptions of the importance of food production' [*Millennium Ecosystem Assessment*, 2005, p. 59]. This perspective is supported by the European Landscape Convention, which defines landscape as 'an area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors' [*Council of Europe*, 2000, p. 2], highlighting human experiences of landscapes. Similarly, local monitoring frameworks, such as, for example, the Swiss Landscape Monitoring Program (LABES) or Landscape Character Assessment (LCA) in England and Scotland, state that it has to be estimated 'how the landscape is perceived and experienced by people' [*Tudor*, 2014, p. 12], and emphasise that 'monitoring of landscape patterns and landscape perception is decisive' [*Kienast et al.*, 2015, p. 136]. The European Landscape Convention requires that member countries identify, assess and monitor landscapes 'taking into account the particular val-

ues assigned to them by the interested parties and the population concerned' [*Council of Europe*, 2000, p. 4].

Moving from over-privileging outstanding landscapes to mapping all landscapes and characterising their distinctive properties through text was an important goal of the LCA framework developed in the 1980s [*Fairclough et al.*, 2018]. This framework is widely adopted within and outside of Europe, and some countries developed their own approaches. These include, for example, the aforementioned LABES in Switzerland, the Landscape Atlases of France and Belgium and the Landscape Observatories of the Netherlands and Catalonia [*Kienast et al.*, 2019]. In the context of perceived landscape properties, these frameworks have been criticised for excluding opinions of people who actually experience landscape by overemphasising expert views [*Butler*, 2016]. This issue is strongly connected to methodological challenges relating to elicitation of human experiences [*Jones and Stenseke*, 2011].

Landscapes are perceived through multiple senses, 'such as smell/ scent, tranquillity, noise, and exposure to the elements (wind and rain for example)', as recognised in official landscape monitoring programs [e.g., *Tudor*, 2014, p. 42]. However, despite this recognition, to date, visual perception is often the only factor assessed [*Kienast et al.*, 2019]. This tendency can be partly explained by the challenges related to capturing multi-sensory human experiences.

Current approaches to the elicitation of public perception of landscapes can be divided into those focusing on in-depth information for small areas and those collecting information over larger territories but having a rather more superficial character. The first group includes different types of interviews and other qualitative approaches [*Caspersen*, 2009; *Clemetsen et al.*, 2011]. The second involves surveys sent to households and analysis of social media data [*Kienast et al.*, 2015; *van Zanten et al.*, 2016; *Tieskens et al.*, 2018]. A public participation geographic information system (PPGIS) can be considered to belong to one group or the other, depending on whether the researchers are in the field together with the participants or if it is hosted online [*Plieninger et al.*, 2013; *Bruns and Stemmer*, 2018]. Interestingly, one of the approaches that originated in the Netherlands – 'Landscape biography' – uses people's stories about the past and historical place names for landscape characterisation. This approach reveals intertwined relationships between people and their historical narratives and landscapes [*Kolen et al.*, 2018].

'Languages are windows on the senses' [*Majid and Levinson*, 2011, p. 7], and can be used as a source of insight into the ways in which different cultures conceptualise landscapes [*Mark and Turk*, 2017] and the ways in which the physical environment is reflected in language [*Regier et al.*, 2016]. Studies in psycholinguistics demonstrated the dominance of references to visual perception in written accounts of different sorts [*Winter et al.*, 2018]. Not surprisingly, European philosophers of the Romantic era prioritised sight over other senses. Judgments

of appearance had to be precise in their wording: terms such as 'picturesque', 'sublime', 'majestic' and 'beautiful' had to be selected carefully in descriptions of views, since they were intended to evoke different sensations in readers [*Donaldson et al.*, 2017]. However, the authors of the Romantic era did not limit themselves to the visual perception of landscapes, by deliberately extending their descriptions with references to sound, noise and silence to enrich readers' understanding [*Agnew*, 2012; *Taylor*, 2018]. Research on historical landscape descriptions emphasises that the inherent ephemerality of sounds makes travel diaries and other historical texts a reliable way to preserve information about soundscapes in a changing environment [*Smith*, 1994; *Taylor et al.*, 2018]. An even more ephemeral sense – smell – and changes in its perception over time and place are also preserved in written accounts [*Corbin*, 1986; *Dann and Jacobsen*, 2003].

Initiatives to digitise texts and make them available to the public undertaken by, for example, the Google Books project, project Gutenberg, British Newspaper Archive, etc. simplify the access to such written accounts [e.g., *Michel et al.*, 2011]. However, their grouping into (potentially) large and coherent sets of texts related to a specific research question – a thematically relevant text corpus – remains an important challenge. The information in such text corpora is available in the form of natural language. The richness of natural language allows us to, for example, identify topics in texts and the sentiments directed towards them [*Drymonas et al.*, 2011; *Jenkins et al.*, 2016]. Temporal references provide information to analyse changes in time [*Adams and Gahegan*, 2016]. Text corpora often contain a plethora of references to locations on earth, and such information is essential to identify spatial patterns [*Gregory et al.*, 2015; *Hu*, 2018]. However, standard automatic text processing methods require adaptation to extract descriptions of first-person landscape perception since they are not developed for the domain of landscape studies, for fine-granular place names and for rural landscapes. Finally, analysing first-person landscape perception available as written narratives requires a critical way of looking at user-generated biases [*boyd and Crawford*, 2011].

1.2 THESIS OVERVIEW

Landscape assessment programs are, we argue, faced with two major challenges: how to take into account multiple voices of those experiencing landscape and how to consider not only visual aspects of landscape perception, but include multi-sensory information. Importantly, the solution has to be reproducible and applicable for different territories.

Insights from different fields of studies suggest that first-person narratives in a form of textual descriptions contain information about perceived landscape properties. Therefore, the overall aim of this work is to explore the added value

of these new sources for landscape assessment. The following four objectives are set out to approach this aim, and the interplay between them and the publications is demonstrated in Figure 1.1.

Objective 1 is to estimate, how reliably landscape properties can be spatially modelled based on textual descriptions.

Objective 2 is to develop a methodological workflow and confirm that textual descriptions contain information about perceived sound experiences and tranquillity and perform spatio-temporal comparison of landscapes based on modern and historical corpora.

Objective 3 is to set out a methodological framework allowing us to automatically collect texts reflecting first-person narratives of landscapes.

Objective 4 is to demonstrate how landscapes can be characterised at different scales using written first-person narratives as sources.

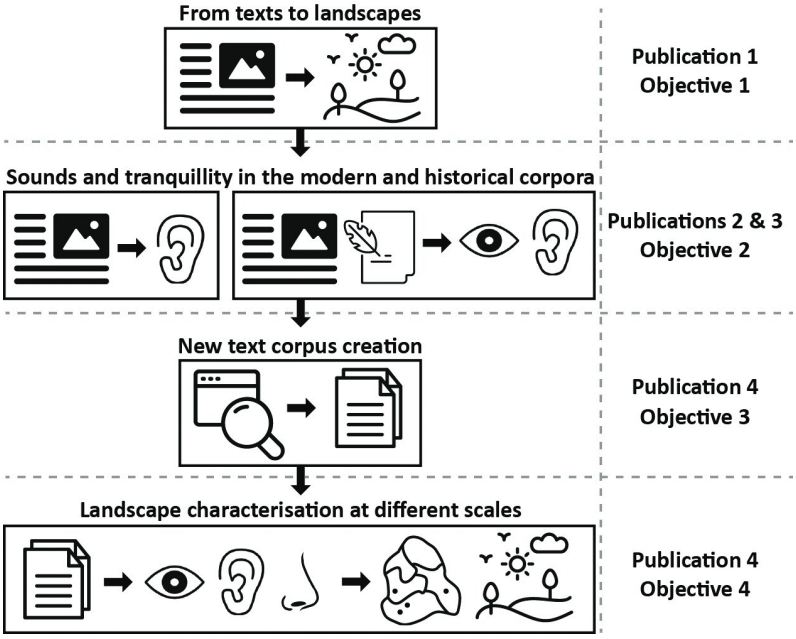


Figure 1.1: Schematic overview of the interrelation between research objectives and publications.

In what follows, we, first, give background information about landscape concepts relevant to this work and outputs of the LCA process, and describe how perceived landscape properties are currently dealt with. We then summarise key methods used to extract and classify information from textual sources. Further, we describe datasets used and created in this work and demonstrate how our methodological workflows can be combined to extract and classify descriptions referring to visual, aural and olfactory perception as well as tranquillity. Finally, we discuss results and limitations of our approach and potential broader implications for landscape research and GIScience.

2

BACKGROUND

2.1 KEY CONCEPTS OF LANDSCAPE PERCEPTION

Cultural and linguistic differences are crucial for international political initiatives, such as the European Landscape Convention [*Burenhult and Levinson*, 2008; *Herlin*, 2016; *Fairclough et al.*, 2018]. The term 'landscape' not only has different meanings in different languages, but is also widely used in everyday language and has changed its meaning over time. This history can shed light on the prevalence of visual perception in landscape research, despite the fact that we also perceive sounds, smells and feel heat and wind. The importance of multi-sensory perception was already emphasised in 1929 in the book *Reine Geographie* (English: Pure Geography) by the Finnish geographer Johannes Gabriel Granö, which was translated to English in 1997 [*Granö*, 1997]. Parallel with the translation of this seminal book, an article appeared by J. Frederick Coeterier, who also argued for including all sensations in the planning process: 'These are the impressions one gets in a landscape from colours, sounds, smells, tastes, humidity, temperature, wind, light and shadow' [*Coeterier*, 1996, p. 38]. In this work, we are not going to explore references to all senses, but we will cover key concepts which allow us to relate language to visual, aural and olfactory perception, as well as tranquillity.

2.1.1 *Landscape and language*

In the opening of the book *Landscape Ecology*, the authors state that the first mention of the term 'landscape' found in written sources, is in the Book of Psalms [*Naveh and Lieberman*, 1994]. The original language of the Book of Psalms is Hebrew, and one has to be careful not to over-interpret the modern translation from Hebrew to English because of the possible differences in meaning from hundreds of years ago. Nonetheless, what is important is that the word used for landscape in Hebrew is etymologically related to 'a beautiful view'.

Indo-Germanic versions of the word 'landscape', including English and Scots, are a compound of land and the suffix –scape (English), -schaft (German), -skap (Swedish), -schap (Dutch), and together they had a political meaning of belonging to a land or to a territory [*Olwig*, 2005]. However, 'landscape' with its connotation of 'scenery' re-entered these languages in the late sixteenth century,

when Dutch painters started to use this term to describe paintings of natural or rural scenery [*Olwig*, 1996]. The Gaelic translation of the word landscape - 'tírdhreach' is a compound of the words tír meaning land and dreach meaning appearance. Drech and derc ('eye, face') both come from Proto-Celtic *derk- 'see' [*Matasovic*, 2009, p. 96]. The Welsh word 'tirlun' corresponds to visual qualities of landscapes and is more related to 'scenery' [*Selman and Swanwick*, 2010]. The word 'tirwedd' is closer to the word landscape, but wedd means appearance, sight, texture, and *tiros- 'land, earth' [*Matasovic*, 2009].

The etymology of the word 'landscape' partly explains the concentration on visual perception for centuries. Another reason is that senses (sight, hearing, smell, touch and taste) are not evenly present in a language. Research on verbs of vision demonstrated that vision dominates in conversation across languages (more than two-thirds of perceptual references), followed by hearing (ranging between 16% and 38%), except for one language – Semai – in which there are more references to olfactory than to aural perception. The prevalence of the rest of the senses varies across languages [*San Roque et al.*, 2015]. Winter et al. (2018) showed that, in the English language not only do references to visual perception dominate in written accounts, but also the lexical variety of visual perception is higher than that of other senses across time and different types of texts (e.g., fiction, academic writing).

2.1.2 *Visual perception and nature conservation in Great Britain*

> 'It will be observed that this country is bounded on the south and east by the sea, which combines beautifully, from many elevated points, with the inland scenery; and, from the bay of Morcamb, the sloping shores and back-ground of distant mountains are seen composing pictures equally distinguished for grandeur and amenity.'
>
> William Wordsworth, 1843,
> *A Description of the Scenery of the Lakes in the North of England*

Visual perception and aesthetics have been subjects of philosophical debates since at least Socrates' time [*Lothian*, 1999]. We are not going to rehash these debates here, but rather give a brief overview of the ideas relevant to our case study region – Great Britain.

Eighteenth-century philosophers actively debated the social, cultural and political implications of aesthetic terminology. In the first half of that century, prominent intellectuals such as Edmund Burke developed influential new definitions of terms such as 'beautiful', 'sublime' and 'majestic' that emphasised contemporary concerns. In his book *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*, written in 1757, Burke suggested that beautiful landscapes were the ones that evoke pleasure, such as smooth rolling hills, while sublime

landscapes, such as mountains, elicited horror [*Olwig*, 2002]. Following this, in the last three decades of the eighteenth century, these debates were continued by aesthetes and landscape artists – such as William Gilpin, Richard Payne Knight and Uvedale Price – who introduced the term 'picturesque' for a scene that could be represented in the style of European landscape paintings [*Brady*, 2003]. Throughout the nineteenth century, literature continued to have an important effect on the ways that British landscapes were perceived and managed; writers like William Wordsworth in the Lake District, the Brontë sisters in Yorkshire, or Thomas Hardy in Dorset influenced – and continue to influence – the way these places are experienced. Sometimes, though, these authors' representations of the landscape followed the picturesque tradition in deliberately overlooking signs of industrialisation [*Donaldson et al.*, 2015; *Herlin*, 2016]. The Victorian era brought with it a certain shift from ignoring local people, as was customary for Gilpin and his contemporaries, to the ideas of John Ruskin, who argued that aesthetics cannot be separated from the quality of life of the local people living in the landscape [*Benson*, 2008]. Ruskin was also an important figure in fighting against industrialisation and its threats to rural landscapes [*Selman and Swanwick*, 2010].

The effects of these philosophical debates on British conservation were profound, since they started numerous movements of nature, scenery and heritage preservation, which led to the creation of a non-government organisation, namely, The National Trust in 1895, co-founded by Octavia Hill, whose family friend was the aforementioned John Ruskin [*Wohl*, 1971] and the subsequent Act to establish the National Trust for Places of Historic Interest and Natural Beauty in 1907 [*Selman and Swanwick*, 2010]. After the first World War, the increase in the urbanised population of industrial workers and the limited availability of land for recreation led to environmental movements and the initiation of conservation groups (e.g., Friends of the Lake District), which played an important role in the establishment of The National Parks and Access to the Countryside Act of 1949. The first four National Parks – the Peak District, the Lake District, Snowdonia and Dartmoor – were designated in 1951 [*Herlin*, 2016; *Friends of the Lake District*, 2019]. An important difference between the National Parks in the USA and those in Great Britain is the degree to which perceived wilderness is to be protected in the former and picturesque landscapes – as appreciated by Gilpin – in the latter [*Olwig*, 2002].

Visual perception was further important in the designation of other protected areas, such as the National Scenic Areas in Scotland and the Areas of Outstanding Natural Beauty in England, Wales and Northern Ireland [*Selman and Swanwick*, 2010]. National Scenic Areas are defined as areas 'of outstanding scenic value in a national context' [*Scottish Natural Heritage*, 2019], whereas Areas of Outstanding Natural Beauty refer to a more multi-sensory appreciation of landscapes; these are areas of 'relative wildness, such as distance from housings or having few roads; relative tranquillity, where natural sounds, such as streams or

birdsong are predominant; natural heritage features, such as distinctive geology or species and habitat' [*Natural England*, 2019].

After several modifications, starting in 1968, the National Park Authority became a government advisory organisation – Natural England – with its counterpart of Scottish Natural Heritage [*Herlin*, 2016]. These organisations have shifted their focus from the best examples of landscapes selected by experts, often based on their scenic qualities, to all landscapes, and developed the framework of Landscape Character Assessment in England and Scotland [*Swanwick and Fairclough*, 2018].

2.1.3 *Landscape Character Assessment framework*

Characterising all landscapes beyond protected areas and documenting their change is at the core of a pioneering landscape monitoring framework – Landscape Character Assessment in England and Scotland (LCA). This framework is widely used in decision-making, e.g., in questions of allocation of areas for wind turbines, transport infrastructure, forest plantations, etc. [*Herlin*, 2016; *Swanwick and Fairclough*, 2018].

LCA includes two stages: desk study and field survey. The process starts with the identification of distinctive patterns by characterising individual elements belonging to 'natural' (e.g., geology, land cover, soils) and 'cultural/ social' groups of factors (historical and current impact of, e.g., land use) (blue and brown colours in Figure 2.1). This first stage is mostly based on desk studies, using, for example, Geographic Information System (GIS) layers and satellite imagery as input. The second stage is concerned with collecting elements from the 'perceptual and aesthetic' group (green colour in Figure 2.1), which are typically assessed by the author of the study in the field, through interviews and observations [*Tudor*, 2014].

The result of this approach is a map with delineated landscape character types and areas of distinctive character. Landscape character types have generic names such as 'estuary and marsh', whereas areas of distinctive character include names of specific places and a list of their distinctive characteristics, as demonstrated in Figure 2.2 for the selected area 29 Wastwater & Wasdale. These characteristics include descriptions such as 'A landscape of contrast...', 'An overwhelming sense of majesty...', 'Predominantly a very tranquil landscape...', 'Major erosion, litter and disturbance impacts from Three Peaks Challenge events...' [*Watkins*, 2008, p. 105].

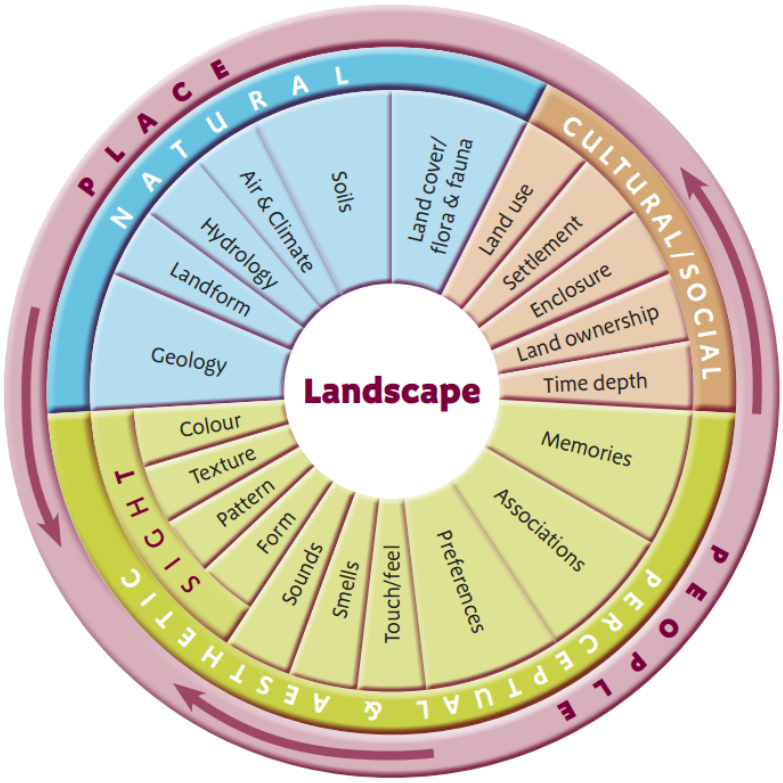
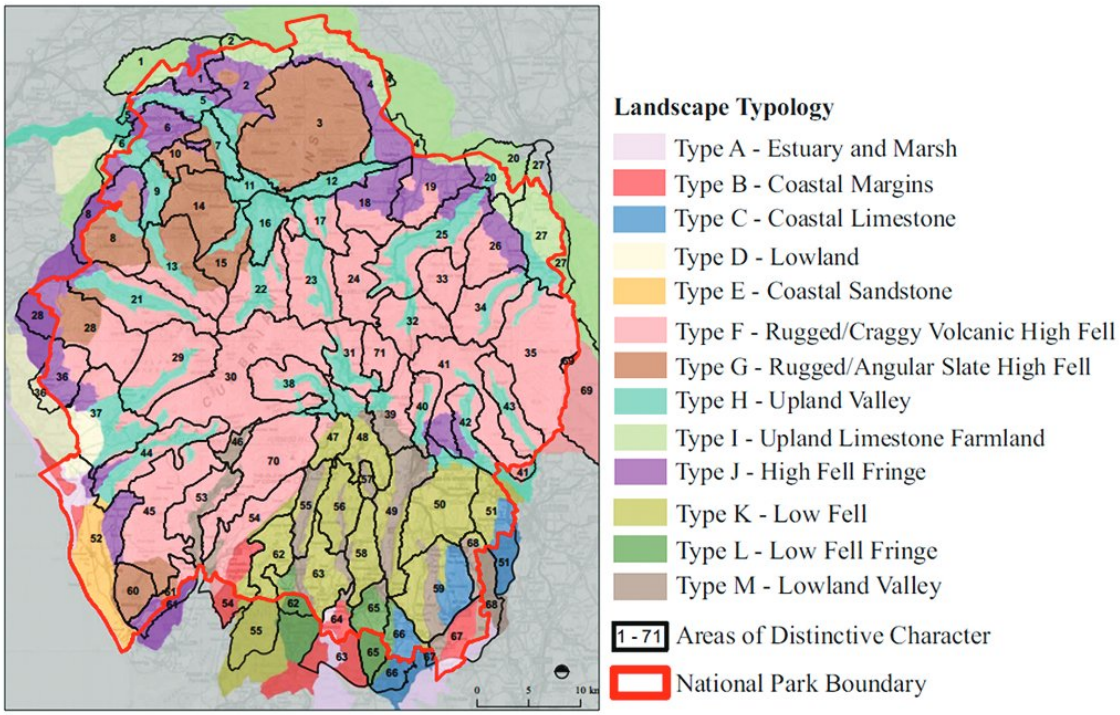


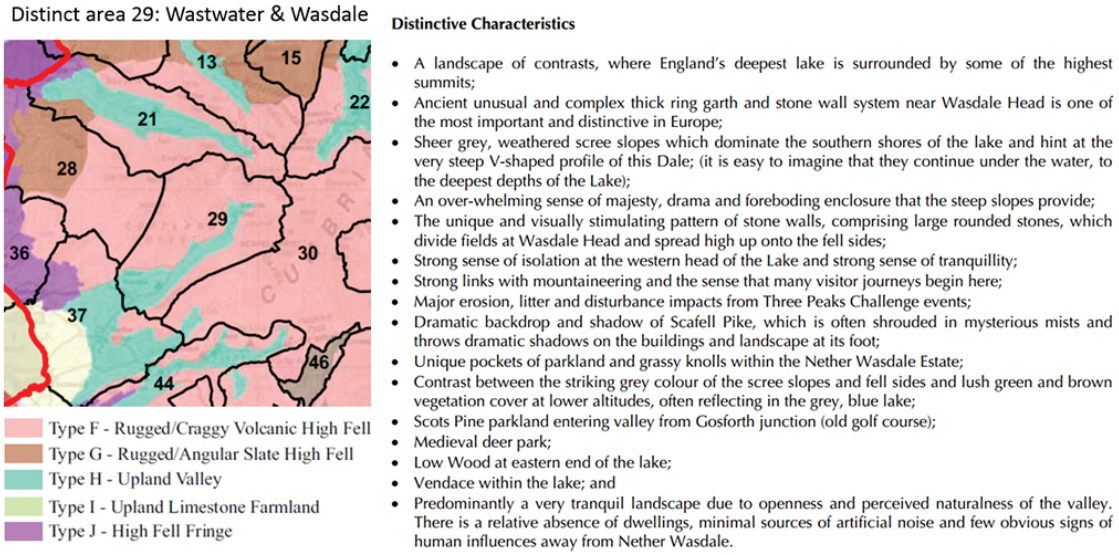
Figure 2.1: Landscape as defined in the Landscape Character Assessment [*Tudor*, 2014].

Additionally, each area has a rich textual description called Local Distinctiveness and Sense of Place. For area 29, this description is 343 words long including the following: 'The valley is popular with climbers and fell walkers, keen to challenge the dramatic surrounding peaks of Scafell, Scafell Pike and Great Gable. The area has a very strong sense of tranquillity due to openness and perceived naturalness of the landscape.' [*Watkins*, 2008, p. 107]. Further characteristics include lists of Landscape Evaluation (e.g., 'Intricate pattern of pasture fields within the valley bottom') and Guidelines for Managing Landscape Change (e.g., 'Conserve and maintain unusual thick ring garth, and other stone walls near Wasdale Head.') [*Watkins*, 2008, p. 107].

One challenge of the LCA approach includes the under-representation of historical landscape processes emphasised in the 'cultural/ social' group as 'time depth'. Historical Landscape Characterisation (HLC) was established in the 1990s to complement LCA with understandings of the history of landscapes and places [*Fairclough and Herring*, 2016]. Other challenges are a disproportionate focus on visual perception in LCA, and both approaches have been criticised for being dominated by expert views and not taking into account multiple voices, for example, in the descriptions related to the 'perceptual and aesthetic' group of landscape elements [*Butler*, 2016; *Swanwick and Fairclough*, 2018]. These twin challenges of integrating multiple views and multiple senses are at the core of this work.



(a) Areas of Distinctive Character and Landscape Character Types



(b) Distinct area 29: Wastwater and Wasdale

Figure 2.2: Lake District National Park LCA report [*Watkins*, 2008].

2.1.4 *Soundscapes*

> 'Coming back to myself, I heard land-birds, starlings, rolled over, looked up at the sky, smelled a sweet smell, some kind of wildflower, thrift maybe.'
>
> Kathleen Jamie, *Sightlines*, 2012

> 'The shoreline forest, as I came back through it, was busy with birdsong.'
>
> Robert Macfarlane, *The Wild Places*, 2008

The term 'soundscape' was coined by R. Murray Schafer alongside the World Soundscape Project in the late 1960s [*Truax*, 1978; *Schafer*, 1993]. In his vision, the concept of soundscape is human-centred, since it is about perception and interpretation of the acoustic environment [*Truax*, 1978]. By contrast, the field of soundscape ecology, which was later renamed *ecoacoustics*, is an ecosystem-centred approach concerned with the ways sounds can be recorded in the field and analysed on the basis of their spectrograms [*Pijanowski et al.*, 2011a, b]. The aim is to study the ways ecosystems are affected by, for example, disturbing anthropogenic sounds. Therefore, it is important to be able to distinguish between, say, air traffic and the calls of birds. However, current methods do not allow clear differentiation between all anthropogenic sounds and sounds produced by animals; for example, birds and human voices are hard to separate, so sample locations must be away from touristic areas [*Pieretti et al.*, 2011].

The difference between human-centred and ecosystem-centred approaches is important in this project, since humans do not perceive all the sounds present in a landscape, but rather hear sounds selectively [*Fisher*, 1999], and 'there's no direct correlation between physical measurements of loudness and perceptions of noise' [*Coates*, 2005, p. 641].

Sound emitters are important in aural perception, since we do not hear abstract sounds, but 'the way things sound' [*Morton*, 2009, p. 40]. Already Granö, in 1929, distinguished between natural and artificial sound emitters [*Granö*, 1997]. Researchers in the field of ecoacoustics proposed a classification of sound emitters into *anthrophony*: sounds produced by people (e.g., traffic noise); *biophony*: sounds of animals (e.g., barking of dogs); and *geophony*: non-biological natural sounds (e.g., wind, waves) [*Krause*, 2008]. Additionally, anthrophony can be divided into human-made sounds and *technophony*, sounds emitted by machines [*Mullet et al.*, 2016].

Why is this differentiation between sound emitters important? Natural and human-generated sounds have opposite connotations in relation to perceived aesthetic quality [*Miller*, 2008]. Middle-distance jet engines emit similar sounds to waterfalls in their pitch and loudness. However, one is perceived as disturbing and the other as 'majestically powerful' [*Fisher*, 1999, p. 28-29]. Unfortunately, it is not that simple; some sound emitters in biophony and geophony

can be considered negative (e.g., howling wolves or thunder), and some anthropogenic sounds – positive (e.g., tolling bells) [*Fisher*, 1999; *Pérez-Martínez et al.*, 2018]. In his book *Walden* written in 1854, Henry David Thoreau goes as far as suggesting that the sound of bells is 'worth importing into the wilderness' [*Thoreau*, 1854, p. 119]. This highlights the importance of analysing the sentiments towards sounds within the context they were expressed in. Such contexts can be provided in rich textual landscape descriptions, as suggested by Prior: 'contemporary nature writing by the likes of Kathleen Jamie and Robert Macfarlane, offers imaginative ways of writing about landscape aesthetics beyond vision' [*Prior*, 2017, p. 12].

Indeed, studies in the humanities have demonstrated that there are numerous references to sound experiences in historical texts [e.g., *Agnew*, 2012]. A surprising example includes 18th century cannon-fire entertainment in the Lake District with the sole aim of listening to the echoes [*Taylor et al.*, 2018]. William Gilpin was one traveller who enjoyed this attraction, yet he also promoted tranquillity and quietness in the Lake District [*Taylor*, 2018].

2.1.5 *Tranquillity*

> 'Next day a brilliant sun spangled the snow and the precipices of Ben a' Bhuird hung bright rose-red above us. How crisp, how bright a world! but, except for the crunch of our own boots on the snow, how silent. Once some grouse fled noiselessly away and we lifted our heads quickly to look for a hunting eagle.'
>
> Nan Shepherd, *The living mountain*, 2011

Tranquillity is a combination of visual and aural perception, recognised as one of the most important aspects of visitor experience in protected and rural areas [*MacFarlane et al.*, 2004; *Miller*, 2008; *Hewlett et al.*, 2017]. The interplay between these two senses is confirmed by experiments, in which respondents found scenes combining positively connotated visual elements (e.g., a stream) with negative sounds (e.g., sounds of a busy park) particularly disturbing [*Carles et al.*, 1999]. Equally, pleasant sounds (e.g., ocean waves) in a disturbing visual context were disliked [*Southworth*, 1969; *Pheasant et al.*, 2008]. Current attempts to classify tranquillity mostly combine GIS layers to produce spatially continuous values on an interval scale from the least to the most tranquil areas [*MacFarlane et al.*, 2004; *Hewlett et al.*, 2017] as demonstrated below in Figure 2.3.

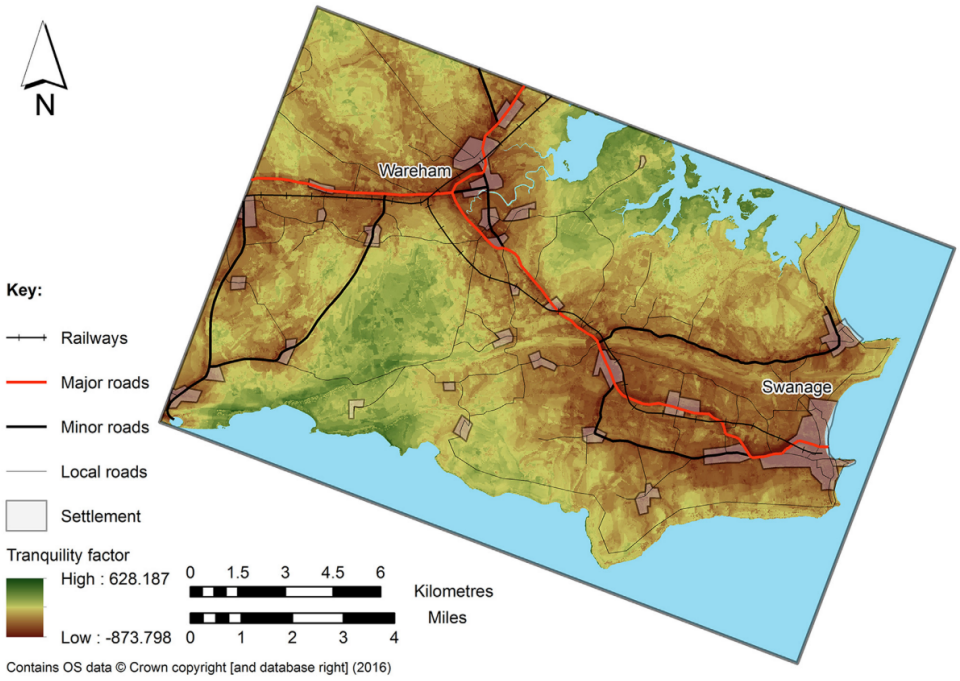


Figure 2.3: Tranquillity map as perceived by visitors of a part of the Dorset Area of Outstanding Natural Beauty [*Hewlett et al.*, 2017].

2.1.6 *Olfactory perception*

> 'The Falls of Measach at Braemore in Ross-shire were certainly a disappointment: too close to a noisy main road and a burger van that filled the deep Corrieshalloch Gorge with the smell of stale fat.'
>
> Anna Pavord, *Landskipping*, 2017

Smellscapes are important in defining the unique character of places, but the English language has a limited vocabulary of odours, making it harder to categorise them. The famous olfactory specialist, Hans Henning classified odours into six primary categories: flowery, foul, fruity, spicy, burnt, resinous [*Majid and Burenhult*, 2014]. Another possibility is classification into positive and negative smell related terms, such as stench and fragrance [*Corbin*, 1986]. However, smells are commonly described by indicating smell emitters, without further classification. We can see in Figure 2.4 that both authors Johannes Gabriel Granö in 1929 and Kate McLean in 2011 chose to indicate the origin of the smell, e.g., the smell of coniferous trees or newly-cut grass.

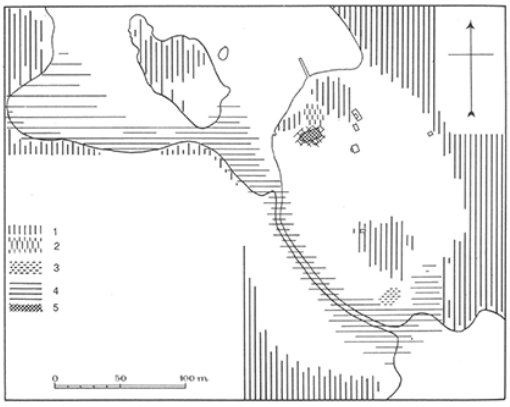
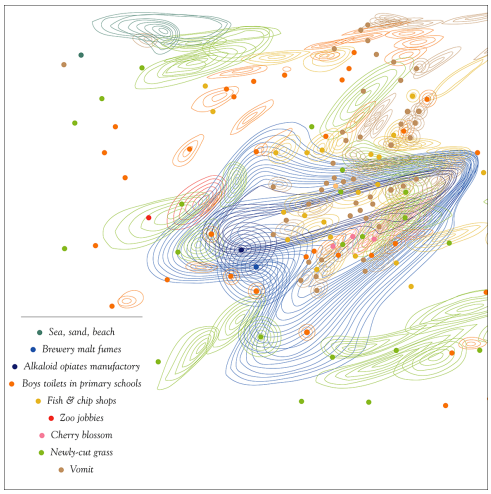


Figure 4.15. Olfactory phenomena of the Valosaari medium during the warmer part of the year.
1 = Smell of coniferous trees; *2* = smell of birch; *3* = smell of bog myrtle (*Myrica gale*); *4* = smell of mud; *5* = variable. Areas with no significant smells are left unhatched.

(a) Valosaari during the warmer part of the year [*Granö*, 1997]



(b) Fragment of Auld Reekie (Edinburgh) on a breezy day [*McLean*, 2011]

Figure 2.4: Maps of smellscapes.

Dann and Jacobsen (2003) claim that, rich context, available in textual sources, is decisive, as smells are hard to describe. To prove their point, they analysed 65 written accounts of fiction distributed in space and time and discovered that writers tend to portray rural landscapes with more favourable smell terms than cities. According to their findings, modern cities (from the Industrial revolution to the post-World War II period) in comparison with pre- and post-modern ones are described using the most negative terms. The authors advocate for distinct smellscapes of touristic destinations and give an example of a successful implementation in Granada, where smells of orange and jasmine in the gardens of Alhambra increase the exoticism of this place and transfer visitors to the days of sultans.

2.2 OVERVIEW OF METHODS USED TO CAPTURE PERCEIVED LANDSCAPE PROPERTIES

An important prerequisite for the study of landscape perception is experiencing it on-site [*Hull and Stewart*, 1992; *Benfield et al.*, 2010; *Dunkel*, 2015], making different types of interviews, surveys, participatory GIS (PPGIS) and social media contributions suitable for eliciting this information. In-depth interviews and focus group discussions provide new insights into the ways people perceive landscapes they have visited and are familiar with. Researchers elicit explicit statements, transcribing and coding interviews, making these approaches time-intensive [*Caspersen*, 2009; *Clemetsen et al.*, 2011]. An additional potential disadvantage could be the absence of geographical coordinates linked to these state-

ments (spatial information is often only implicit and on the coarse granularity level).

In Bieling et al. (2014), the authors claim that another form of interviews – free-listing – can be carried out in very little time. However, similarly to on-site surveys, interviewers have to be in the field with the people who are experiencing the landscape. This makes the methodology relatively expensive in time and resources, leading to small test areas. Moreover, both methods can potentially intrude into the experiences of people in wild nature [*Webb et al.*, 1966; *Taylor et al.*, 1995]. Additional difficulties include the importance of the initial question, in the case of free-listing [*Wartmann et al.*, 2015], and questionnaire design, in the case of surveys. Free-listing methods in landscape studies can be directly spatially linked to the location where the question was posed, whereas on-site surveys can potentially collect information about other locations and larger regions with an implicit spatial component. A solution to this challenge is provided by PPGIS methods, in which explicit spatial information is an inherent property [*Brown and Reed*, 2009; *Plieninger et al.*, 2013]. Nevertheless, PPGIS and the methods listed above suffer from lack of participation, despite a variety of engagement activities and advertisements [*Bruns and Stemmer*, 2018; *Bubalo et al.*, 2019].

Surveys sent to households in a given region are more suitable for larger territories, however, usually only the opinions of locals experiencing landscapes are taken into account, leaving out visitors to the area [*Kienast et al.*, 2015]. Analysis of social media contributions (e.g., Flickr, Instagram) is another method that is also suitable for larger territories and overcomes the previous limitation, as tourists often upload their landscape experiences on social media platforms [*Gao et al.*, 2017]. Interestingly, people who agree to participate in visitor surveys and people who share their experiences online overlap only by 38% [*Heikinheimo et al.*, 2017], with the younger population being more likely to be covered by social media [*Chen et al.*, 2018]. However, exact demographic information is lacking for social media contributions, as opposed to carefully designed interviews and surveys.

Spatial information in social media can be both implicit and explicit, since some of the images have coordinates and some contain information about their location in the tags (e.g., tag 'Matterhorn'). In addition to their position, social media contributions also contain other types of information: user name, time stamp when the image was taken and time stamp of the upload, the photograph itself and descriptions or tags. Perceived landscape properties can be linked to the position of taken pictures [*Tenerelli et al.*, 2016] or to the number of individuals taking pictures per area unit [*Casalegno et al.*, 2013; *Gliozzo et al.*, 2016].

Other studies have analysed not only the position, but also the content of the images (which objects and colours are present on a photograph) [*Richards and Friess*, 2015; *Seresinhe et al.*, 2017]. Despite recent advances in machine learning

that have made it possible to automatically recognise objects in photographs, only objects that were attributed in the training data can be recognised (e.g., only 102 outdoor scene attributes in the SUN database[1]). Tags selected by the photographer herself can provide understandings of what was noticed or considered important and potentially include qualitative assessments of landscapes (e.g., tag 'overcrowded') [*Edwardes and Purves*, 2007]. However, the motivations for tagging are complex and may include social recognition, attracting attention or organising photo collections [*Marlow et al.*, 2006].

Time stamps uploaded with social media allow us to monitor changes over time or across seasons [*Dunkel*, 2015; *Lim et al.*, 2018]; and since the data is collected regardless of whether it is used for research or not, we can 'go back in time' and add another area of interest. Therefore, if, for example, a survey was not conducted for an area of interest at some specific moment in time, it is still possible to elicit this information from social media. Combining locations and tags can allow modelling landscape preference across large extents, for example, the whole of Europe [*van Zanten et al.*, 2016], providing a good basis for visual perception elicitation, but photographs and tags alone cannot easily convey information about other senses. Despite attempts to include sound experiences [*Quercia and Schifanella*, 2015] and olfactory perception [*Aiello et al.*, 2016], tags often only reflect potential sources (e.g., if there is a tag 'church', we can potentially hear a peal of bells).

Daume et al. (2014) suggested using other types of information available online in the context of identifying unanticipated threats in forest monitoring, namely online travel diaries, where 'Author's profile information, personalised and often opiniated content is common' [*Daume et al.*, 2014, p. 11]. At the same time, Bieling (2014) demonstrated the richness of information available in 14 short stories collected in a 'short-story contest' and categorised this information into groups useful for the framework of Cultural Ecosystem Services (e.g., references to aesthetics). Wartmann et al. (2018) compared information from 50 texts from online travel diaries, Flickr tags and on-site free-listing interviews, all collected for the same 10 locations. Their results demonstrated that online travel diaries are richer in perceptual landscape aspects than free lists and Flickr tags.

These results suggest that descriptions of first-person landscape perception available in travel reports and other written accounts have a potential for going beyond visual perception and expert views and overcoming the challenges related to the elicitation of public perception of landscapes, since they contain explicit statements about landscape perception, information about the author and a time stamp, allowing us to collect landscape perception from the past. Spatial information is present in such descriptions at a fine-granular level; however, it is implicit, meaning that the descriptions must be linked to space computationally [*Purves and Derungs*, 2015].

1 http://groups.csail.mit.edu/vision/SUN/

Key messages:

- Insights from the fields of psycholinguistics, ethnophysiography and humanities suggest that references to visual perception will be dominant in written first-person narratives, but that they will also contain information about other senses.
- Over-privileging of the most beautiful landscapes through preconceived notions of landscape appearance, established in the philosophical debates of the 18th and 19th centuries, was shifted to allow inclusion of all kinds of landscapes in the framework of Landscape Character Assessment (LCA).
- Each area of distinctive character, as used in LCA, has associated rich textual descriptions created by experts, whose views are criticised for being dominant.
- Human-centred approaches to aural perception focus on sounds perceived by people, as opposed to all sounds present in the environment, which are, like smells, typically classified based on the emitter of the sound or smell. Tranquillity, however, is most commonly classified on a continuous scale from the least to the most tranquil areas.
- Some studies in landscape research suggest using online travel diaries, which are claimed to have information about the author's profile and 'opiniated' content.

2.3 METHODS OF TEXT ANALYSIS

In the following section, we describe the methodological tools used in this work. These tools can be divided into several groups, which are summarised in Table 2.2 with their main properties and exemplary works, with the focus on geographic analysis. The important terminology appearing throughout this chapter is summarised in the box 'relevant terminology'.

To illustrate the tools, we will use the same sentence written by Thomas Gray in his letter to Wharton in October 1769, describing landscapes around Ullswater in the English Lake District:

It is soon again interrupted by the roots of Helvellyn, a lofty & very rugged mountain

Relevant terminology:

- **User-generated content:** images, videos, text or audio content contributed by users on online platforms.
- **Web scraping:** extraction of data from online resources.
- **Unstructured texts:** texts that lack a data-model and metadata, and cannot be easily indexed or organised into a database.
- **Text corpora:** large and coherent sets of texts.
- **Document:** the basic entity of a corpus. A document can be one newspaper article, all travel blog entries of one user, etc.
- **Tokens:** the finest units of analysis reported in this work, e.g., words, punctuation.
- **Stopwords:** a list of frequent words (e.g., 'the') which are commonly filtered out before any processing.
- **Lemmatisation:** determining the base form – lemma – of a word, e.g., 'tree' and 'trees' both have the lemma 'tree'.
- **Part of speech tagging:** labelling each word as noun, adjective, etc.
- **Word sense disambiguation:** identifying in which sense a word is used in a sentence (e.g., a 'sound' as auditory sensation or a 'sound' as a geomorphological form).
- **Supervised and unsupervised text classification:** the difference is the fixed set of classes that are labelled in the training data in the former and grouping of texts based on similarities in the latter.
- **Macro- and micro-analysis:** terms coined in the context of digital humanities by Matthew L. Jokers (2013), referring to automatic processing of big textual collections in the former and careful reading of individual works in the latter.
- **Feature vector:** a vector of features that, in our case, represents characteristics of texts.
- **Training and test data:** sets of data used to fit the parameters and to assess performance of a model.
- **Sentiment analysis:** classifying opinions expressed in texts.
- **Georeferencing:** association of information (e.g., texts or sentences) with locations on the earth.
- **Gazetteer:** a dictionary of place names which contains names, their coordinates and types.
- **Toponym recognition:** identification of place names in texts.
- **Toponym resolution:** assignment of unique coordinates to place names.
- **Geographic scope of documents:** set of locations representing the content of a document or a spatial footprint of a document.
- **Precision:** the proportion of correctly selected answers to all answers.
- **Recall:** the proportion of correctly selected answers to all possible correct answers.

2.3.1 *Creating text corpora*

The idea of compiling collections of texts is not new; in the fields of descriptive linguistics and literary studies, such systematic collections, for example, of the complete works of one author, have been known at least since 1950s. Already in the 1960s, the Brown Corpus (Brown University Standard Corpus of Present-Day American English)[2] and in the 1980s its British English counterpart, the Lancaster-Oslo-Bergen corpus[3], were created [*Pustejovsky and Stubbs*, 2012]. However, recent developments in the digitisation of books and newspaper archives [*Michel et al.*, 2011; *Nicholson*, 2012], and a tremendous increase in the volumes and variety of data available online [*Ginsberg et al.*, 2009; *Hu*, 2018], opened up new possibilities in other fields of research; for example, to analyse regional differences in the ways people write about space [*Xu et al.*, 2014], to discover broad topics written about large geographical areas [*Adams and McKenzie*, 2013], or to demonstrate spatio-temporal changes in mentions of concepts such as steam and electricity in Great Britain [*Lansdall-Welfare et al.*, 2017].

Many general corpora have already been created and made available for research. These include corpora designed to be representative of a language, such as the Brown Corpus mentioned above, and corpora of the whole web collected during a particular time period, such as Common Crawl[4] and ClueWeb12[5]. Many thematical corpora have also been compiled, such as a corpus of news (e.g., Reuters Corpus [*Rose et al.*, 2002]) or corpora of travel writing of the same region (e.g., the Swiss Alpine Club corpus Text + Berg [*Volk et al.*, 2009], the historical Corpus of Lake District Writing [*Donaldson et al.*, 2015]). Another type of existing corpora are those created and made available for the training and testing of different algorithms, e.g., a corpus of Tweets GeoCorpora [*Wallgrün et al.*, 2018].

To ensure thematic, spatial and temporal relevance to a specific research question, it is possible to build a corpus by extracting data from online resources, so-called web-scraping. One possibility is to scrape the whole context of specific webpages, e.g., English Wikipedia [*Adams and Gahegan*, 2016] or specific travel blog webpages [*Drymonas et al.*, 2011]. To cover specific spatial regions, a set of search terms can be based on place names or postal codes present in the area of interest [e.g., *Davies*, 2013; *Xu et al.*, 2014; *Wartmann et al.*, 2018]. In studies of spatial language, a typical approach is to use patterns as search seeds, such as 'hotels in <toponym>' [*Jones et al.*, 2008] or '<toponym> near <toponym>' [*Derungs and Purves*, 2016a].

2 https://www.sketchengine.eu/brown-corpus/
3 http://www.helsinki.fi/varieng/CoRD/corpora/LOB/
4 http://commoncrawl.org/
5 https://lemurproject.org/clueweb12/

2.3.2 *Pre-processing*

Texts available in such corpora are in the form of unstructured text. This is written content, which, as opposed to structured data, lacks a data-model and metadata, and cannot be easily indexed or organised into a database [*Hu*, 2018].

Depending on the computational task, a text can be divided into documents, sentences or individual tokens. While document chunking is a conceptual decision, division into sentences and tokens is a computational task, which is less trivial for a machine than it is for a human. In the case of sentences, one challenge is dealing with ambiguous punctuation (e.g., 'e.g.'), and in the case of tokens, among others, with hyphenation related to the width of the page. Historical texts are more prone to these challenges since they can have idiosyncratic punctuation, case and hyphenation [*Butler et al.*, 2017]. When texts are divided into tokens, the next common step is to filter out so-called stopwords (e.g., 'and', 'a', 'the'). Despite being criticised for some machine learning tasks [*Agarwal and Yu*, 2009], this step is powerful in reducing noise in tasks like topic modelling and identifying common co-occurrences. Additionally, lemmatisation, stemming and normalisation of words to their lower cases are common pre-processing steps. Lemmatisation and stemming derive base forms by removing inflectional endings or affixes. The difference is that, by lemmatisation, the base form is the root word or lemma, whereas in stemming, it is the root stem. Lemmas are words that can be found in a dictionary, whereas root stems may not always be actual words [*Manning and Schutze*, 1999]. For example, the Lancaster stemmer[6] changes Gray's words *It is soon again interrupted by the roots of Helvellyn, a lofty & very rugged mountain* to:

it is soon again interrupt by the root of helvellyn , a lofty & very rug mountain

2.3.3 *Ways of dealing with unstructured text*

Unstructured text may be handled in different ways: as a bag-of-words (independent tokens), n-grams (sequences of tokens), applying rules or patterns, and searching for dependencies (Table 2.1). For the first two options (bag-of-words and n-grams), the part of speech (e.g., adjective, noun) of the words is optional but commonly used. For rules and patterns, part of speech is useful, depending on the task. For example, for the pattern '<toponym> near <toponym>', part of speech labels are not crucial [*Derungs and Purves*, 2016a], whereas clearly important for the pattern 'noun + adjective'. For dependencies, part of speech labelling is essential [*Manning and Schutze*, 1999].

6 https://github.com/words/lancaster-stemmer

Table 2.1: Ways of handling unstructured text.

bag-of-words	The order of the words and their dependencies do not matter, but their frequency and co-occurrence with each other play an important role. [It; is; soon; again; interrupted; by; the; roots; of; Helvellyn; ,; a; lofty; &; very; rugged; mountain]
n-grams	An n-gram is a given sequence of text; for example, bigrams are sequences of two tokens. [It is; is soon; soon again; again interrupted; interrupted by; by the; the roots; roots of; of Helvellyn; Helvellyn ,; , a; a lofty; lofty &; & very; very rugged; rugged mountain]
patterns detection	An example of a pattern could be 'adjective + noun' as, for example, 'rugged mountain' [*Kisilevich et al.*, 2010].
dependencies	Defining the syntactic dependencies between the words is a complex task, despite the availability of libraries [*Hall et al.*, 2011]. However, they are powerful in, for example, identifying that Helvellyn is lofty, that Helvellyn is very rugged, and that Helvellyn is a mountain (Figure 2.5).

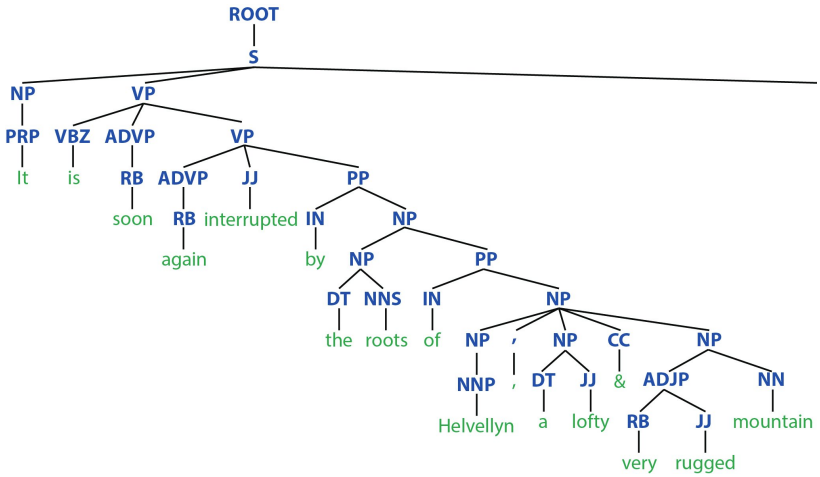
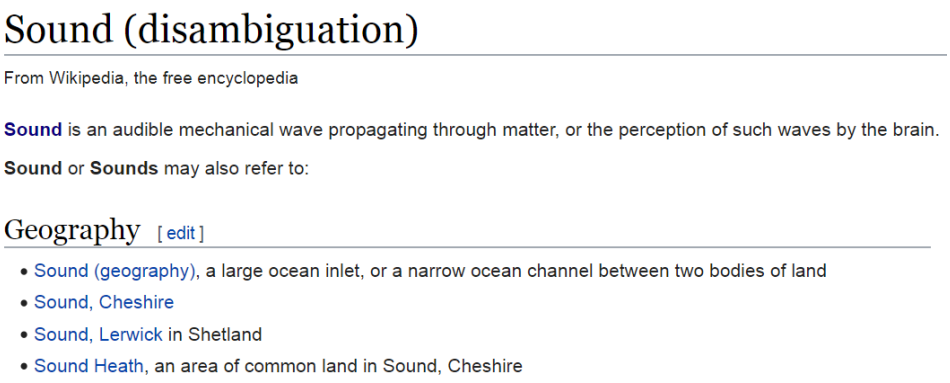


Figure 2.5: Illustration of our example processed by the Stanford dependency parser. Blue letters stand for types of dependencies and in their lowest hierarchical level for different part of speech labels, e.g., 'Helvellyn, a lofty & very rugged mountain' is a noun phrase (NP), 'very rugged' is an adjective phrase (ADJP), 'very' is an adverb (RB) and 'rugged' is an adjective (JJ).

Importantly, in part of speech tagging, most available off-the-shelf tools return not only labels, such as nouns, verbs, adjectives, and adverbs, but also more complex ones, such as modal verbs, proper nouns, verbs in past tense, etc. Note that, in our example, the past participle verb 'interrupted' is wrongly labelled as adjective (blue capital letters JJ in Figure 2.5).

Ambiguity is an inherent property of natural language and, consequently, of texts. Many words have counterparts that are spelled in the same way, but have a different meaning, such as polysemic words that are semantically related (e.g., metaphorical use of the word 'roots' in *the roots of Helvellyn*) or homonyms, semantically unrelated words, e.g., a 'sound' as auditory sensation or a 'sound' as a geomorphological form (Figure 2.6). As people do, machines can imply meaning depending on the context of the word in question. The process of finding the right meaning is called word sense disambiguation. One widely used algorithm to resolve such ambiguities is the one invented by Lesk. This algorithm uses hypernyms (categories) extracted from WordNet [*Fellbaum*, 1998] and compares the definition of a target word with definitions of context words [*Lesk*, 1986; *Vasilescu et al.*, 2004]. This algorithm works well for verbs and nouns; however, adjectives and adverbs are not arranged in a hierarchy in WordNet, therefore, hypernyms cannot be used. A simple strategy for such cases is to control for parts of speech. For example, to keep the word 'still' when it is an adjective and used as 'motionless', and discard it if it is an adverb used as 'yet'.


Sound (disambiguation)

From Wikipedia, the free encyclopedia

Sound is an audible mechanical wave propagating through matter, or the perception of such waves by the brain.

Sound or **Sounds** may also refer to:

Geography [edit]

- Sound (geography), a large ocean inlet, or a narrow ocean channel between two bodies of land
- Sound, Cheshire
- Sound, Lerwick in Shetland
- Sound Heath, an area of common land in Sound, Cheshire

Figure 2.6: Example of word sense disambiguation as approached by Wikipedia.

2.3.4 *Extracting relevant text snippets*

Commonly, candidate documents or sentences have to be extracted from a corpus before their further classification or other forms of content exploration can be performed; this process can be called creation of a sub-corpus. One common starting point is a list of keywords that is often combined with detection of patterns and syntax dependencies. For example, to extract potential descriptions of

fictive motion, Egorova et al. (2018) used a list of nouns collected from mountaineering glossaries and searched for it in a corpus using a pattern: 'nouns from the list followed immediately by a verb' (e.g., 'the ridge runs') and a pattern with a syntax dependency: 'nouns from the list linked to a verb by a determiner' (e.g.,'a ridge that runs').

Another possibility is to extend a list of initial keywords by calculating which words co-occur with them more often than by chance [*Gablasova et al.*, 2017]. A special case of this approach is to limit the co-occurring words to place names. For example, Murrieta-Flores et al. (2015) used this technique to relate different diseases to locations in Great Britain.

In order to take into account meanings of words it is possible to use word embeddings, where words are represented as vectors with nearby vectors being more similar to each other in meanings [e.g., *Pennington et al.*, 2014]. For example, McGregor and McGillivray (2018) used word embeddings to extend a list of smell-related keywords to search in historical medical reports. However, the main limitation of word embeddings is that different meanings of a word are conflated into a single vector representation. If this technique is used only to extend the set of keywords, this limitation could be overcome by adding an additional step of word sense disambiguation, as described in Section 2.3.3.

2.3.5 *Classification*

If we wish to differentiate between, for example, different types of activities people perform in a city as reflected in textual data, we can perform unsupervised classification by grouping texts based on their statistical similarities with further labeling of the emerging classes. Another possibility is supervised classification, where classes are defined in advance, for example, for types of activities these could be 'education', 'entertainment', 'recreation', 'sports', etc. To accomplish this, annotated training data has to be available or created prior to the classification.

If the annotated data is newly created, a set of clear rules has to be established. The data is annotated by at least two annotators, and then inter-annotator agreement is calculated using a statistical measure, typically a Kappa Statistic [*Landis and Koch*, 1977; *Pustejovsky and Stubbs*, 2012]. If inter-annotator agreement is not sufficient, the rules have to be refined.

As input many machine learning algorithms use a feature vector, rather than raw unstructured text. Texts can be represented as frequency- or binary vectors, where the presence (or frequency) of a word, n-gram, pattern (e.g., 'noun + adjective') or syntax dependency is checked in every document. Another way to create such vector-representations is to use as input normalised term frequency tf-idf (term frequency - inverse document frequency), which lowers the values

of words used commonly across the whole corpus [*Manning and Schutze*, 1999]. Another possibility is to represent the whole paragraph or document as one vector, as suggested by Le and Mikolov (2014). All the possibilities listed in Table 2.1 are commonly used as features for machine learning classification, with uni- and bigrams, their part of speech (and the aforementioned tf-idf) being the most common ones [*Liu*, 2012].

A feature vector representation of Gray's sentence, based on frequency of nouns, adverbs and adjectives as identified by a part of speech tagger and length of the sentence is $[n_{nouns}, n_{adverbs}, n_{adjectives}, n_{tokens}]$ or [3, 3, 3, 17], since each of these part of speech categories appears three times in our sentence (Figure 2.5), and it contains 17 tokens.

These feature vectors are further used as input for a classifier (e.g., naïve Bayes, SVM) [*Pang et al.*, 2002]. One commonly used classifier is random forest, since it can also be used for regression problems; does not require any assumptions with respect to data distribution; is robust to noise in training data, and the effect of overfitting is seldom seen, as random forest creates random subsets of the feature vectors and builds smaller trees using these subsets. Lastly, it requires less training data than deep learning [*Criminisi et al.*, 2011].

Sentiment analysis is an example of a classification task, which can be performed using machine learning techniques as described above or using a lexicon. A typical problem of these approaches is the lack of annotated training data or of a domain-specific lexicon [*Choi and Cardie*, 2009]. General-purpose lexicons, such as the Opinion Lexicon [*Hu and Liu*, 2004], exclude some domain-specific polarities. From our example, only the word 'interrupt' is in the negative list of the lexicon and the synonym of the word 'rugged' - 'rough'. Words such as 'mountain' and 'lofty' are not present. However, using pretrained word embeddings [e.g., *Pennington et al.*, 2014], it is possible to extend sentiment values from the Opinion Lexicon to other words [*Iyyer et al.*, 2015]. One time-efficient way to build a domain-specific lexicon is to use additional clues to assign polarity to words or sentences. These can be pros-and-cons tables, common in reviews of technical equipment, or stars, common in reviews of movies [*Kaji and Kitsuregawa*, 2007; *Lu et al.*, 2011]. However, existing lexicons are often either too general or domain-specific in a rather narrow domain. Typically, sentiments are classified either into three classes: positive, neutral and negative [*Liu*, 2012] or according to emotions (e.g., anger, fear, joy, sadness) [*Resch et al.*, 2016; *Lim et al.*, 2018].

2.3.6 *Assigning texts to space*

Before starting with the analysis of spatial patterns emerging from a text, toponyms mentioned in a text must be identified (toponym recognition), associated with unique coordinates (toponym resolution) and grouped into a spatial footprint (identification of geographical document scope). Toponym recognition and resolution steps include disambiguation and are often done by comparing tokens with existing digital gazetteers and lexicons [*Jones et al.*, 2008; *Purves et al.*, 2018]. A digital gazetteer is a dictionary of place names, which typically contains the name, coordinates of the location, type of place and country, as demonstrated in Figure 2.7 [*Hill*, 2009].

Potential challenges in toponym recognition include structural ambiguity (e.g., does 'Geneva' refer to a city in Switzerland or is it a part of the toponym 'Lake Geneva'?), referent class ambiguity (e.g., does 'Gary' refer to a man's name or is it a settlement in Indiana, USA?) and reference ambiguity, where one place has several names (e.g., do 'Matterhorn' and 'Cervino' refer to the same mountain?) [*Purves et al.*, 2018].

The toponym resolution step most often includes referent ambiguity, which means that the same toponym can refer to more than one geographic place, e.g., Helvellyn in Grenada, Australia, UK or South Africa, Figure 2.7 [*Amitay et al.*, 2004; *Overell and Rüger*, 2008]. Moncla et al. (2014) introduced a new type of ambiguity – unreferenced toponyms ambiguity. It refers to the incompleteness of gazetteers and is very relevant to this work since it is important in georeferencing fine-grained toponyms [*Acheson et al.*, 2017].

helvellyn | all countries | search | [advanced search]

5 records found for "helvellyn"

	Name	Country	Feature class	Latitude	Longitude
1	Helvellyn	Grenada, Saint Patrick	populated place	N 12° 13' 26''	W 61° 37' 52''
2	Helvellyn Rocks	Australia, Queensland	rocks	S 20° 50' 0''	E 149° 18' 0''
3	Helvellyn Helvellyn,he er wei lin feng,هلوِلين,赫爾維林峰	United Kingdom, England Cumbria > Eden District > Patterdale	mountain elevation 950m	N 54° 31' 42''	W 3° 1' 10''
4	Helvellyn Helvellyn	South Africa, Eastern Cape Joe Gqabi District Municipality > Senqu	mountain	S 30° 42' 0''	E 27° 16' 0''
5	Helvellyn	South Africa, Eastern Cape Joe Gqabi District Municipality > Senqu	farm	S 30° 41' 51''	E 27° 17' 54''

Figure 2.7: Results of the query 'Helvellyn' in the GeoNames gazetteer illustrating referent ambiguity [*GeoNames*, 2019].

Depending on the spatial granularity of texts (e.g., coarse-granular news articles versus fine-granular travel diaries or hiking blogs), the methods of disambiguating toponyms vary. The majority of the developed methods have been applied to coarse-granularity texts (e.g., information about size of population or hier-

archy of administrative units), which we are not going to describe in detail, we rather list heuristics that can be used for fine-granular texts as well:

- Location of the 'source': ambiguous toponyms tend to be closer to the location of the source, which can be, for example, home-town of the writer [*Buscaldi and Magnini*, 2010].
- Geographic contribution: toponyms that appear in the same document tend to refer to locations that are close to each other distance-wise [e.g., *Leidner et al.*, 2003; *Moncla et al.*, 2014].
- Text contribution: spatial inference, if an ambiguous toponym refers to a location, for example, 'south of unambiguous one', that might resolve the ambiguity [*Leidner and Lieberman*, 2011].
- Combination of geographic and text contributions: toponyms and other words, which are close to the ambiguous toponyms in the document (toponym sequence or text distance) can either indicate place types [*Martins et al.*, 2008; *Buscaldi and Magnini*, 2010] or be helpful for the disambiguation [*Smith and Crane*, 2001].
- One sense per discourse: if a toponym is mentioned multiple times in the document, it always refers to the same location [*Leidner et al.*, 2003; *Amitay et al.*, 2004; *Martins et al.*, 2008].

A set of the most representative locations in a document can be considered a geographic document scope [*Amitay et al.*, 2004; *Monteiro et al.*, 2016]. However, in this work we are more interested in the ways a document can be represented by its spatial footprint. After the steps of toponym recognition and resolution, a document is represented by a set of point locations, since geometric information in most of the gazetteers is stored in the form of points. Furthermore, irrelevant points have to be filtered out or the most important ones identified. The step from a set of points to a spatial footprint can be done by assigning them to different types of grid cells [*Derungs and Purves*, 2016b; *Hobel et al.*, 2016], by using kernel density estimation [*Hollenstein and Purves*, 2010], clustering [*Ahern et al.*, 2007; *Gao et al.*, 2017] or by convex hull [*Wartmann et al.*, 2018].

Table 2.2 summarises methodological tools described in Sections 2.3.1-2.3.6 and gives an overview of exemplary works using these techniques for geographical applications. Most of the examples perform a combination of tasks, e.g., extraction and classification, however, we have described them in the table only once for a selected task.

Table 2.2: Exemplary works of geographic applications using methods from Section 2.3.

Geographical application	Exemplary study, data & methods
2.3.1 Creating text corpora	
spatial language	[*Jones et al.*, 2008], querying the web with a pattern
local place knowledge	[*Davies*, 2013], querying the web with place names
thematic regions	[*Adams and McKenzie*, 2013] scraping predefined travel blogs
regional differences in spatial language	[*Xu et al.*, 2014], querying the web with postal codes and the keyword 'directions'
sense of place	[*Wartmann et al.*, 2018], querying the web with place names and the keyword 'we'
2.3.4 Extracting relevant text snippets	
natural features	[*Derungs and Purves*, 2014], SAC yearbooks[7], manual annotation, querying keywords, spatial tf-idf
references to particular diseases	[*Murrieta-Flores et al.*, 2015], Histpop corpus[8], co-occurrences of place names with keywords
references to place classes	[*Ballatore and Adams*, 2015], scraped travelblog.org, automatically created list of keywords
aesthetic terminology in travel writing	[*Donaldson et al.*, 2017], CLDW[9], co-occurrences of place names with a list of four thematic keywords
fictive motion	[*Egorova et al.*, 2018], The Alpine Journal[10], querying patterns and syntactic dependencies

7 A historical corpus of the Swiss Alpine Club yearbooks documenting 150 years of Alpine mountaineering, a part of the Text + Berg corpus
8 A historical corpus spanning from 1801 to 1937 and covering statistics and textual descriptions of the births, deaths and marriages of the British population
9 A historical Corpus of Lake District Writing
10 The Alpine Journal 1969-2008 is a part of the Text + Berg corpus

Table 2.2: Exemplary works of geographic applications using methods from Section 2.3.

Geographical application	Exemplary study, data & methods
2.3.5 Classification	
sentiment analysis	[*Drymonas et al.*, 2011], four scraped travel blogs, classification based on existing annotated data
urban soundscapes	[*Aiello et al.*, 2016] Flickr tags, lexicon-based, co-occurrence network
emotions related to urban places	[*Resch et al.*, 2016] Tweets, manual annotation, graph-based algorithm
emotions related to green spaces	[*Lim et al.*, 2018] Tweets, lexicon-based, frequency of each term
landscape values	[*Chen et al.*, 2018] Instagram captions, semi-inductive content analysis
2.3.6 Assigning texts to space	
vague place names	[*Hollenstein and Purves*, 2010] Flickr tags, kernel density estimation
thematic regions	[*Jenkins et al.*, 2016] Tweets, Wikipedia, spatial clustering approach
spatial folksonomies[11]	[*Derungs and Purves*, 2016b], SAC yearbooks, hiking blog Hikr, grid-based approach
vague cognitive regions	[*Hobel et al.*, 2016] Trip Advisor, hexagonal grid approach
vague cognitive regions	[*Gao et al.*, 2017], Flickr, Tweets, Instagram, Wikipedia, scraped travel blog, delineation of clusters

11 Spatial folksonomies are defined by the authors as 'a tuple linking a vocabulary of terms through authors and resources to locations' [*Derungs and Purves*, 2016b, p. 61]

2.3.7 *Evaluation measures*

The most common evaluation measures used in natural language processing are precision and recall. Precision is a measure of the proportion of the selected correct answers to the total of selected answers (true positives (green) divided by the sum of true positives and false positives (green and blue areas) in Figure 2.8). Recall is a measure of the proportion of selected correct answers to the total of correct answers (true positives (green) divided by the sum of true positives and false negatives (green and yellow areas) in Figure 2.8) [*Purves et al.*, 2018].

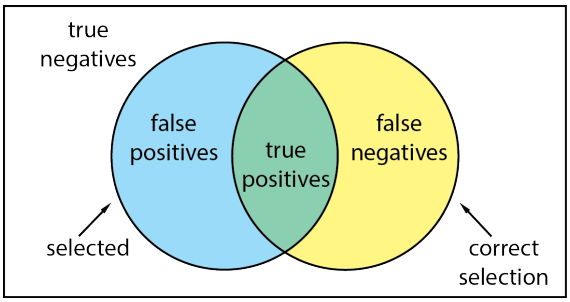


Figure 2.8: A diagram demonstrating the correct selection (yellow) and the selected set (blue), and notions of true and false positives and true and false negatives, adapted from [*Manning and Schutze*, 1999].

2.3.8 *User-generated biases*

Michel et al. open their seminal paper with the line 'We constructed a corpus of digitized texts containing about 4% of all books ever printed' (2011, p. 176); this constitutes millions of books. There are several consequences of these numbers: researchers analysing such data might assume that, since the volumes of data are tremendous, content biases introduced by individuals are minimal [*Purves et al.*, 2011] and spatial coverage is evenly-distributed. Both of these assumptions are false, almost all media information has a geographical bias towards the Global North [*Graham et al.*, 2014], and is prone to participation inequality, where only small number of users contributes the majority of the content [*Nielsen*, 2006; *Purves et al.*, 2011; *Haklay*, 2016]. Additionally, the 4% of books included are not a carefully-selected random sample, but rather, books available in particular libraries [*Lansdall-Welfare et al.*, 2017].

Olteanu et al. (2019) recently published a comprehensive overview of a variety of biases present in social media. Of course, the first step in dealing with such biases is to recognise them, but it is even more important to recognise which of these biases are actually relevant to research question. Since this problem is not new, there are accepted ways of dealing with it. In landscape value modelling, researchers moved from counting numbers of contributions to numbers of indi-

vidual contributors per grid cell or other spatial units, to ensure that individual contributors do not bias the results [*Gliozzo et al.*, 2016; *van Zanten et al.*, 2016]. From the content point of view, one possibility is to see how often a given word is used by different contributors, depending on how prolific they are [*Purves et al.*, 2011]. Figure 2.9 demonstrates such an approach by comparing two Flickr tags: 'london' and 'innercity', revealing that 'innercity' is not a commonly-used tag, as opposed to 'london', and keeping it in the analysis can bias the results [*Hollenstein and Purves*, 2010].

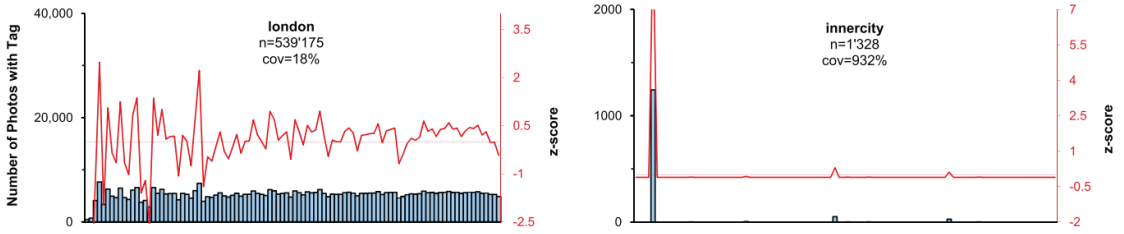


Figure 2.9: Term profiles for Flickr tags 'london' and 'innercity', demonstrating that 'london' is ubiquitously used by different users, but 'innercity' only by a small group of prolific users. The histogram is sorted by placing the most prolific users on the left [*Hollenstein and Purves*, 2010].

A bigger challenge is to demonstrate, whose opinions such data actually represents, since there is a lack of detailed demographics about contributors to different types of user-generated content. In the case of social media, some insights are already known; for example, young adults are more likely to contribute to social media platforms [*Toivonen et al.*, 2019], and there is no significant difference between women and men for some platforms [*Hausmann et al.*, 2018]. As described in section 2.2, people contributing to social media are not the same as those agreeing to take part in visitor surveys [*Heikinheimo et al.*, 2017], making the combination of different approaches crucial.

2.4 RESEARCH GAP AND RESEARCH QUESTIONS

To implement requirements of different international environmental agreements, in particular the European Landscape Convention, a reproducible workflow is necessary that allows the capture of multi-sensory personal experiences and perception. However, this task is not trivial; a variety of different methods can be used to understand and collect perceived landscape properties. Unsurprisingly, each of the methods has its advantages and disadvantages. The most relevant disadvantages of existing approaches include small study areas, time intensity, inflexibility, absence of the experiences of some groups of people, and conclusions about perceived properties based on implicit data. In the field of landscape research, the idea of using different forms of written texts is gain-

ing weight, and the presence of references to perceived landscape properties is confirmed by the way written accounts of landscape descriptions were and are being created, at least since the Romantic era. The automatic processing of texts has great potential for being a scaleable and reproducible method, which considers on-site first-person experiences. Therefore, the overall aim of this work is to demonstrate how information extracted from written landscape descriptions can complement current methods of landscape assessment.

To approach this aim, we set out to investigate the following research questions:

RQ.1 How can a corpus be created in order to capture first-person perception of rural landscapes?
RQ.2 How reliably can we model spatial variation in landscape properties using text corpora?
RQ.3 Which perceived landscape properties of rural landscapes can be extracted from written landscape descriptions and how do these properties vary?
RQ.4 To what degree is it possible to compare landscapes based on landscape descriptions?
RQ.5 How can the connection between landscape properties captured from written landscape descriptions and landscape assessment be established?

3

FROM TEXT SOURCES TO PERCEIVED LANDSCAPE PROPERTIES

To move from text sources to perceived landscape properties, we, first, give important insights into the study area (Great Britain) and case study area (the English Lake District). Further, we describe the existing datasets used in this project and the way we created an additional one – a text corpus of first-person landscape perception in the Lake District. In what follows, we demonstrate how language can be used to spatially model landscape scenicness, an important step in legitimising our further exploration of textual descriptions as a source of landscape perception information. We then set out the methodological framework that allows us to automatically collect texts reflecting individual experiences of landscape in relation to visual, aural and olfactory perception as well as tranquillity. Since our aim is to explore the potential of these extracted texts for landscape monitoring and assessment, we discuss the ways landscapes can be compared and characterised at different levels of granularity, from the coarse level of Great Britain to the fine-grained level of individual landscape elements. As we apply a mixture of macro- and micro-analysis, where appropriate we use footnotes to quote the original description. We report on intrinsic measures of performance when describing each method, before summarising the results of an extrinsic validity through experts' discussion at the end of the chapter.

3.1 STUDY AND CASE STUDY AREAS

> 'I am always glad to see Staveley; it is a place I dearly love to think of – the first mountain village that I came to with William when we first began our pilgrimage together.'
>
> Dorothy Wordsworth, *Journals of Dorothy Wordsworth*, 1897

The long tradition of landscape assessment programs and nature conservation in Great Britain makes this region an exemplary location to test the potential of written descriptions as a source of information on first-person landscape perception. An additional advantage is that it is an English-speaking region. Natural language processing methods in English outperform those in other languages not only in quality, but also in the variety and availability of lexicons and annotated data.

Our case study region is the English Lake District, an area in the North-West of England (Figure 3.1). As described in Section 2.1.2, the Lake District region was one of the first areas in Great Britain to be designated as a National Park, in 1951 [*Herlin*, 2016]. This was possible through the help of the charity organisation, Friends of the Lake District, which was established in 1934 with the aim of protecting and promoting the landscape of the Lake District [*Friends of the Lake District*, 2019].

The total area of the Lake District National Park is 2362 square km, and the majority of land is privately owned (58.8%). The National Trust owns about one quarter (24.8%), followed by the water company United Utilities (6.8%), Forest Enterprise (5.6%), Lake District National Park Authority (3.8%) and the Ministry of Defence (0.2%) [*Watkins*, 2008]. The famous children's author, Beatrix Potter lived in the Lake District and was an important conservationist of the local farming traditions, including preservation of Herdwick sheep native to the Lake District (Figure 3.3b). She acquired farms, ensuring their survival and granted the land she owned (about 16 square km) to the National Trust [*Watkins*, 2008; *National Trust*, 2019a]. In June 2019, the National Trust acquired one more piece of land in the Lake District – Brackenthwaite Hows – a location from which J.M.W. Turner painted his watercolour Crummock Water in 1797, making it the first site bought by the National Trust exclusively for its panorama [*Pidd*, 2019]. In 2017 the Lake District received UNESCO World Heritage Status, motivated by the landscape's beauty and tranquillity, farming traditions and inspiration it provided not only to artists and writers (e.g., Samuel Taylor Coleridge, Dorothy and William Wordsworth, J.M.W. Turner), but also to influential ideas about relationships between people and landscape based on the emotional response to it [*Nomination*, 2017].

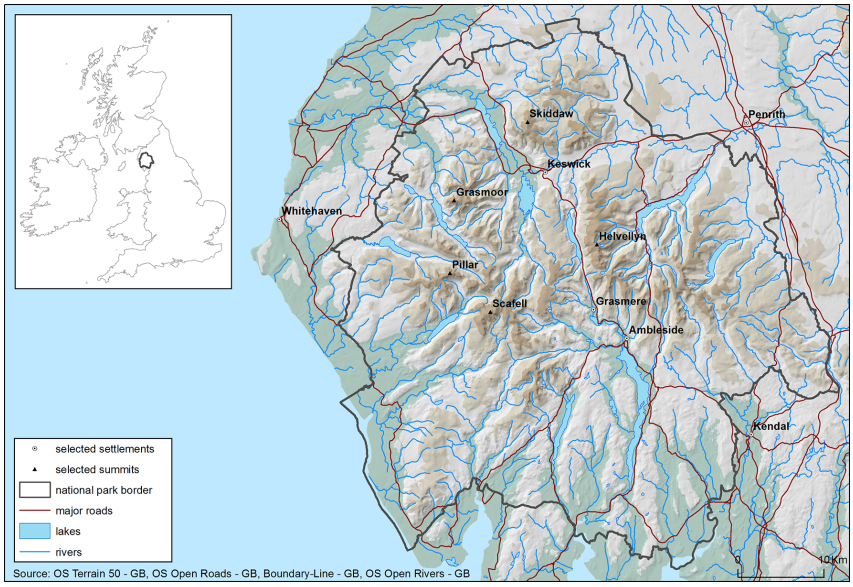


Figure 3.1: Our study area: The Lake District in the North-West of England. Figure from Chesnokova et al. (2019) – Publication 3.

The Lake District includes England's highest mountain – Scafell Pike (978 m), often climbed in the context of the Three Peaks Challenge (highest points of England, Wales and Scotland, see one of the descriptions in Figure 2.2) and sixteen major lakes. Smaller water bodies in the region are called *tarns* and are located high up in the mountains, known as *fells* (Figure 3.3).

In parallel with the 'discovery' of the Lake District by British travellers interested in aesthetic appreciation in 18th-19th century, this area became a centre of different types of industry, including quarrying of slate, limestone, and granite; many of the quarries are still visible in the landscape today [*Watkins*, 2008; *Donaldson et al.*, 2015]. As we described in Section 2.1.2, 19th century writers such as William Wordsworth deliberately ignored these signs of industrialisation. William Wordsworth was the one to write to the editor of the *Morning Post* to oppose the extension of the railway to Windermere in 1844, since it would bring countless number of visitors and ruin the peace of the Lake District [*Taylor*, 2018]. In 1847, the railway was opened and 'mass tourism' started to emerge, also as a result of increases in wages and free time [*Watkins*, 2008]. Today the Lake District National Park is visited by more than 16 million people yearly, almost twice as many as Yorkshire Dales and the Peak District [*National Parks UK*, 2019]. Visitors to the Lakes are involved in many outdoor activities offered by the region, such as fell climbing and wild swimming, often in the form of the so-called 'hill-bagging' and 'tarn-bagging', since the Lake District is home to various lists put together by writers and bloggers [*Crocker and Jackson*, 2001]. Some examples of such lists include:

- 214 peaks, described by Alfred Wainwright in his seven-volume *Pictorial Guide to the Lakeland Fells* and referred to as 'Wainwrights' (Figure 3.2)
- 110 summits, listed in *The Outlying Fells of Lakeland* by Alfred Wainwright
- 541 hills over 1000 feet, listed by Bill Birkett in his *Complete Lakeland Fells*
- 332 Lake District tarns, listed by John and Anne Nuttall
- Tarns, listed by W Heaton Cooper in *The Tarns of Lakeland*

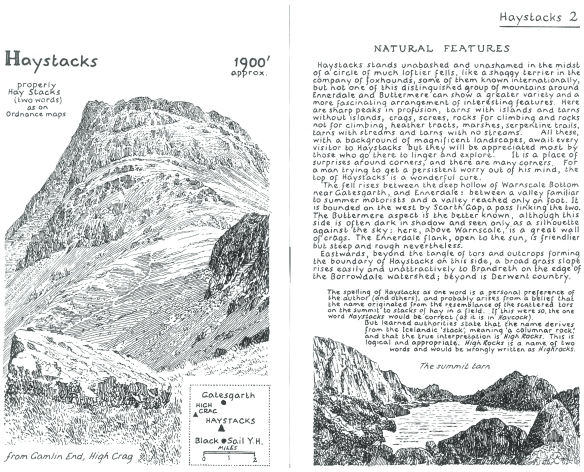


Figure 3.2: Fragment of the *Pictorial Guide to the Lakeland Fells* [*Wainwright*, 1966].

3.2 DATASETS USED IN THIS WORK

Great Britain and the Lake District are also well-suited study areas since they are the locations of unique datasets containing rich texts, such as the crowdsourcing projects, *Geograph British Isles* and *ScenicOrNot* and a collection of 80 texts put together in the historical *Corpus of Lake District Writing* (CLDW) (Table 3.4).

The Geograph British Isles[1] project was launched in 2005 with the aim of collecting representative photographs and textual descriptions of every square kilometre of Great Britain and Ireland. It is created in a game format to motivate contributors with a points-system, so that they submit more than one photograph per square kilometre and cover all the grid cells. This strategy has led to a collection of more than 6 million georeferenced photographs contributed by more than 13000 authors, and to geographical coverage not strongly biased to urban areas. In addition to the photographs themselves and their descriptions, the following properties of the contributions are known: user ID and name, title, tags, image class (e.g., lake, stone circle), date and coordinates. The dataset is available under a Creative Commons Licence.

(a) Buttermere and Crummock Water

(b) Herdwick sheep

(c) Derwent Water

(d) Angle Tarn

Figure 3.3: Typical landscapes of the Lake District National Park (Photos: Olga Koblet).

1 http://www.geograph.org.uk/

ScenicOrNot[2] is a project collecting ratings of scenicness of Geograph photographs. Users can rate randomly shown photographs from 1 (not scenic) to 10 (very scenic) without seeing the description to it since it could influence the results of the rating [*Hodgson and Thayer*, 1980]. The combination of Geograph photographs and their description with the ScenicOrNot votes is demonstrated in Table 3.1. The dataset is available under an Open Database Licence.

Table 3.1: Examples of Geograph contributions with the ScenicOrNot votes.

Photo		
Description	Farm buildings in early mist just outside of Hodnet.	It was reputedly here that King Arthur received his legendary sword, Excalibur.
URL	www.geograph.org.uk/photo/3585	www.geograph.org.uk/photo/9190975
Author	Andy and Hilary	Roger Geach
Date	15 April, 2005	28 July, 2008
Votes	5, 5, 7, 6, 5, 6	5, 3, 6, 1, 5, 8

As suggested in the description of user-generated biases (2.3.8), the majority of contributions in Geograph were submitted by a small number of prolific users. For example, in the Lake District a single user contributed around 11000 photographs of which ca. 850 were rated in the ScenicOrNot project. Detailed demographic information about Geograph contributors is not available, but based on a survey from the project initiators, the authors are most likely to be males over 50. As opposed to Geograph, where photographs and descriptions are contributed by authors who actually visited the landscape, the ScenicOrNot project is purely internet-based and no information about the voters is available. Both projects have been successfully used in studies relating human well-being and scenicness [*Seresinhe et al.*, 2015], modelling scenicness [*Jeawak et al.*, 2017] and detecting cultural ecosystem services [*Gliozzo et al.*, 2016].

The CLDW is a georeferenced collection of novels, poetry, non-fictional essays, letters and travel writing about the Lake District and its surroundings spanning from 1622 to 1900 and including not only famous, but also lesser-known writers,

2 http://scenicornot.datasciencelab.co.uk/

such as Joseph Budworth, Catherine Hutton and Harriet Martineau [*Donaldson et al.*, 2017; *Murrieta-Flores et al.*, 2017]. In this work we selected only non-fictional texts as the most similar to the Geograph corpus and the corpus we describe in the following section.

Additionally, for sentiment analysis tasks we used a general 'Opinion Lexicon'[3] containing around 6800 English words categorised as either positive or negative [*Hu and Liu*, 2004].

3.3 CREATING A TEXT CORPUS

The uniqueness of the project Geograph British Isles makes it necessary to create another corpus of texts describing our case study region the English Lake District, since we would like to demonstrate the transferability of findings to other regions. Geograph descriptions are undoubtedly first-person perception narratives. To ensure this in our new text corpus, we came up with a set of annotation rules in respect to three classes: first-person perception of landscapes (e.g., 'A thankfully short unpleasant section through conifers, no sound, no vegetation and hardly any light.'[4]), landscape descriptions which do not describe individual experiences (e.g., 'Routes starting from Skiddaw Forrest in the east are also quieter, giving the walker a sense of being in the wilderness.'[5]), and irrelevant descriptions (e.g., official parish information, weather forecasts). The distinction between the first two classes was introduced both to separate idealised views of landscapes typical in advertisements of accommodations and guided walks, and to exclude generalised views over seasons and years, typical in descriptions related to navigation information.

Examples of first-person landscape descriptions include the following characteristics:

- explicit descriptions of perception ('the heather smells lovely')
- events that have already happened as opposed to anticipated ones
- descriptions using verbs of motion in combination with personal pronouns 'I' or 'we' ('we went to...'; 'I walked 12 miles')
- potentially contain references to time ('today'; 'this lovely morning')
- potentially contain descriptions of weather ('it was still raining')

Landscape descriptions which do not describe individual experiences:

- present a consistent use of passive voice ('it can be done by both car and on foot')
- contain imperatives ('keep on the road')

3 https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
4 https://www.andrewswalks.co.uk/lordsseatgroup.html
5 http://english-lake-district.info/skiddaw/skiddaw.html

- describe anticipated events ('next week we go to the magnificent Aira Force waterfall')
- contain information to help navigation ('the views to your left (east) are breath-taking')
- indoor descriptions

In the pilot study, we selected 10 neighbouring areas of distinctive character (from a total of 71) in the Lake District, with each referred to using a set of characteristic toponyms (e.g., Wastwater & Wasdale in Figure 2.2). These toponyms were used as search terms while querying the web through the BootCaT toolkit [*Baroni and Bernardini*, 2004]. Based on the returned URLs from the first five areas, we created a list of terms that should not be a part of the URLs (e.g., 'hotel'). Further, we manually annotated all 641 retrieved texts into the three classes described above. Table 3.2 gives an overview of the relations between types of search terms and the number of retrieved texts. We observe that the type 'hill' has the highest number of both first-person landscape descriptions and documents with landscape descriptions not related to individual experiences, which are very similar to each other, whereas water bodies have a high number of retrieved documents especially in the class of descriptions not related to individual experiences. Interestingly, not every hill is as important as every other. The hills used as search terms are all listed in the seven-volume book written by Alfred Wainwright in the fifties, where he describes how he climbed 214 fells (Section 3.1, Figure 3.2). The relatively low total number of texts is explained by BootCaT limitations including the maximum of 100 returned pages, which are then filtered for the English language and other rules defined by BootCaT, e.g., average lengths of sentences of the webpage segment, its relative position, etc [*Baroni and Bernardini*, 2004]. This has also influenced the number of not-relevant documents; many of them were filtered so strongly that they would contain only one sentence.

The main results of the pilot study were:

- 641 annotated documents,
- list of terms that cannot be a part of the URLs (e.g., 'hotel'),
- identified advantage of the 'Wainwrights' list of 214 summits as search terms, and
- identified limitations of the BootCaT toolkit.

Based on these results we developed a workflow that involves six major steps. The first is to select relevant search terms, considering geographical and thematic coverage. To ensure different types of authors in our final set of texts, in addition to the Wainwrights list, we added the top 150 suggestions by TripAdvisor on sights and landmarks in the Lake District, as used in [*Richards and Tunçer*, 2018]. To increase the number of the first-person perception descriptions returned, we added the personal pronoun 'I' to the search terms. In this step,

Table 3.2: Search terms (can be several per area of distinct character), their type according to the Ordnance Survey gazetteer and the number of texts per class.

Search term	Type	Total	First-person perception	Not individual experiences
Bassenthwaite	settlement	37	1	9
Bassenthwaite Lake	water	48	2	21
Blencathra	hill	56	14	16
Blindcrake	settlement	27	1	2
Broom Fell	hill	62	24	7
Caldbeck	settlement	36	0	9
Embleton	settlement	38	0	3
Kirk Fell	hill	61	18	18
Ling Fell	hill	52	24	12
Lorton Vale	other	46	2	13
Loweswater	water, settlement	42	1	12
Mungrisdale	settlement	32	7	4
Skiddaw	hill	43	7	14
Uldale	settlement	25	2	2
total		641	103	150

since we used names as search terms, we had to deal with reference ambiguity (e.g., Blencathra is also known as Saddleback and Hallsfell Top) and referent class ambiguity (e.g., Pillar and Sail are both names of peaks and common words in English) (Section 2.3.6). These two ambiguities were handled by querying with all known names for the location in the former and by querying with both names alone and, additionally using the search term 'wainwright' in the latter.

For the second step, we queried with Bing Web Search API[6], since it does not limit the number of returned URLs, and as opposed to BootCaT, it only delivers the list of the URLs without further scraping (extracting data from websites). On the one hand, it requires performing an additional step in the workflow, but on the other hand, we are not influenced by the strong filtering of content that occurs in BootCaT. After the potential URLs are retrieved, we filter duplicates and those that are most likely to be irrelevant, using the list created in the pilot study.

6 https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

In the third step we scrape[7] visible textual content from the webpages present in our list of the URLs, excluding texts in headers, footers, sidebars, and comments. The scraping is performed in accordance with robots.txt files of each webpage, which are defined by the webpage author and have the possibility of forbidding scraping. Having performed this step, we have a collection of documents.

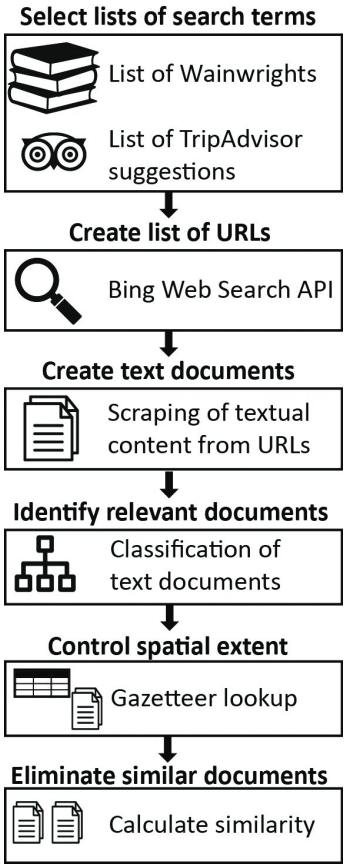


Figure 3.4: Creation of the first-person landscape perception corpus.

The fourth step is the classification of all retrieved documents into three classes described above in order to identify texts describing first-person landscape perception. The annotated data was created in the pilot study, and we used random forest classification[8], trained and tested on half of the data, to classify the newly retrieved documents. We achieved a precision of 0.84 assessed on 641 texts by using the following features: the 250 most frequent words; presence of selected personal pronouns; and the 50 adjectives and nouns common in each of three classes separately.

In the fifth step, we ensured that the documents classified as first-person landscape perception are actually about the Lake District by a simple toponym recognition step using the Ordnance Survey gazetteer limited to the Lake District and applying fuzzy matching through the Levenshtein distance, since spelling and capitalisation in texts and official sources often differ.

Lastly, we identified and removed similar documents using 80% string similarity between the descriptions. This step is important since we do not want to bias our corpus by adding the same description several times.

Having performed these steps, we created a corpus of 6870 spatially and thematically relevant documents, with the characteristics summarised in Table 3.3. Characteristics of text corpora used in this work are summarised in Table 3.4 and external datasets in Table 3.5.

7 https://scrapy.org/
8 https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Table 3.3: Characteristics of the corpus of first-person perception in the Lake District.

Characteristic	TripAdvisor	Wainwright
Number of search terms	92	233
Initial number of extracted texts	13110	24150
Relevant texts in the Lake District	961	5909
Average sentences per text	81	104
Average words per text	1277	1120

Table 3.4: Textual corpora used in the project and their description.

Name	Description
Geograph UK http://www.geograph.org.uk	More than 6 million georeferenced photographs of every square kilometre of Great Britain and Ireland, contributed by more than 13000 authors, with relatively rich text descriptions.
Geograph Lake District http://www.geograph.org.uk	Around 91000 photographs contributed by more than 1200 authors are within the Lake District region. Of these, about 65000 photographs include descriptions, which total 1.4 million words.
CLDW [*Donaldson et al.*, 2017]	Georeferenced corpus of 80 texts and over 1.5 million words written in the 17th-19th centuries about the Lake District. The corpus contains more than 36000 georeferenced locations in the UK, of which 70% are in the Lake District. Non-fictional version of the corpus contains 61 texts written by 55 authors and comprises 1.3 million words.
Text corpus of first-person perception in the Lake District	The corpus contains about 7000 documents in the Lake District comprising almost 8 million words with the average number of sentences per text being 92 and the average words per text – 1198.

Table 3.5: External data sources for the project and their description.

Name	Description
ScenicOrNot http://scenicornot.datasciencelab.co.uk	More than 220000 Geograph photographs were rated at least 3 times on the ordinal scale from 1 (not scenic) to 10 (very scenic). Information about the voters is not available.
Opinion Lexicon [*Hu and Liu*, 2004]	A general lexicon containing 2006 positive words and 4783 negative words including miss-spellings.
LCA polygons	Boundaries of the 71 areas of distinctive character in the Lake District provided by the Lake District National Park Authority.

3.4 LANGUAGE AS A SCENICNESS PREDICTOR

Scholarship from different fields increasingly advocates using data associated with natural language as a bottom-up source of information for landscape preference estimation, particularly in relation to visual perception. To corroborate the validity of this assertion, we tested if landscape *scenicness* could be predicted purely from textual descriptions associated with photographs contributed in the project Geograph. As described in Section 3.2 and demonstrated in Table 3.1, over 220000 Geograph photographs are rated online in the frame of the project ScenicOrNot. Therefore, by assigning these votes to the Geograph descriptions and using machine learning we can estimate the predictive ability of language.

First, we had to perform a set of basic pre-processing steps to generate a feature vector described in detail in Section 2.3. To these, we added an additional step, specific to the task of modelling: filtering out toponyms from the descriptions, since they can bias the model. To do so, we used the official local gazetteer Ordnance Survey for a simple gazetteer look-up method. The photographs in the Geograph project are georeferenced, making it possible to select only toponyms which are within 5 km radius from the photograph's location. This allows us to account for referent class ambiguity, since 'Flat' and 'Green Hill' can be toponyms, but we want to keep these words if the description is not made in the vicinity of these toponyms. Further, we combine each description with the votes of scenicness of its associated photograph. After these pre-processing steps we have, for each photograph, a combination of its ID, normalised tokens (unigrams and bigrams), their part of speech and votes (e.g., '8; ridge; NN;

8,3,8,5,8,6,8', where NN is noun). If identical tokens have different parts of speech, they are stored separately (e.g., '84; traffic; JJ; 2,2,3,5,2,3' and '10194; traffic; NN; 9,1,1,2,6,1,2,2,2,1,2,2,5', where JJ is adjective and NN is noun).

Second, in supervised models it is necessary to have training and test datasets. Informed by potential biases introduced by participation inequality described in Section 2.3.8, we divide our descriptions on training and test datasets using two configurations, taking account of the spatial autocorrelation in user contributions: fully random and user-dependent random, where we allow descriptions of individual users to be only in one half.

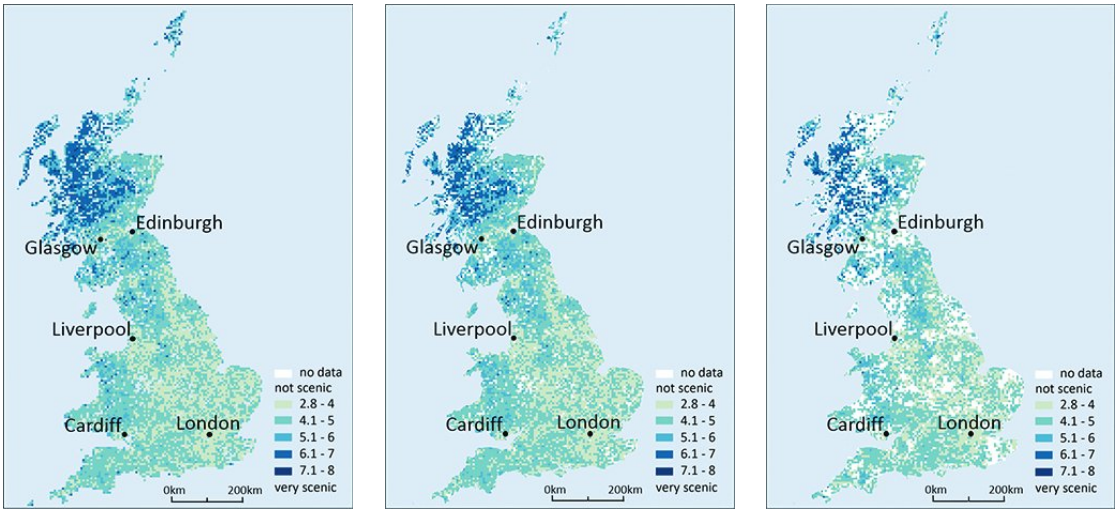
The final step is to create a spatially contiguous model of scenicness for which we decided to use 5 km grid cells. We have additionally tested 2.5 km and 10 km grid cells and found that 5 km was a good compromise between reduced model performance due to limited number of descriptions per grid cell and smoothed variation that is easier to predict.

To model scenicness we used a random forest regression, and the highest explained variance – 52.4% in the case of fully random and 52.0% in the case of user-dependent random configurations – was achieved using the 800 most frequent unigrams, presence of adjectives from 'Landscape Adjective Checklist' by Craik [*Nasar*, 1992] and weighting according to spatial tf-idf [*Rattenbury and Naaman*, 2009]. These results are comparable with traditional approaches using interviews and participatory methods [*Palmer*, 2004], land cover data [*Stadler et al.*, 2011] and social media [*van Zanten et al.*, 2016], demonstrating that textual descriptions are feasible to use in studies of landscape perception.

Figure 3.5 shows both spatial patterns of predicted scenicness and their original distribution. The patterns are very similar to each other, with Scotland to the north-west of Edinburgh and Glasgow being the most scenic, but what is striking here is the large number of grid cells with no value in white colour for the user-dependent random configuration. This demonstrates the effects of participation inequality, since cells with no value mean that all photographs within these cells were taken by a single user. However, high similarity in the values of explained variance suggest that the restriction of descriptions of individual users to be either in training or in test datasets is unnecessary.

3.5 PERCEIVED LANDSCAPE PROPERTIES

Motivated by the results of landscape scenicness modelling, we continue our exploration of natural language for landscape perception, focusing on visual, aural and olfactory perception and tranquillity. Since we are interested in characterising landscapes, we start by identifying useful classes of each phenomenon, followed by the extraction of relevant text snippets – our sub-corpus – and its further classification. The final step is to link these classified descriptions to



(a) Original distribution (b) Fully random configuration (c) User-dependent configuration

Figure 3.5: Maps of the scenicness prediction results. Figure adapted from Chesnokova et al. (2017) – Publication 1.

space at different levels of granularity, described in Section 3.5.5. We repeat this workflow for every type of landscape perception with different inputs of corpora, classes, keywords and methods of classification (Figure 3.6). In the classification step, we used three approaches: lexicon-based, machine learning and we annotated our data by hand where the complexity of descriptions was too high (for historical texts) or the volume of available training data was too low (for olfactory descriptions).

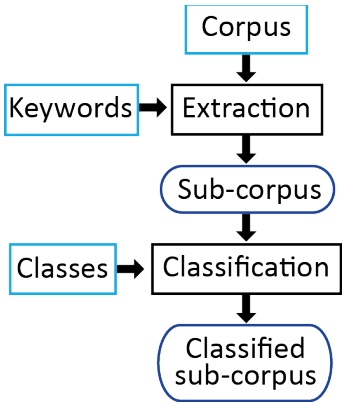


Figure 3.6: Overall workflow used to extract and classify descriptions of landscape perception. Black colour squared boxes indicate processing steps, dark-blue boxes with rounded corners – output, and light-blue squared boxes – external sources.

3.5.1 *Visual perception*

To classify visual perception, we limited ourselves to identifying descriptions related to the most scenic and least attractive landscapes for the whole of the UK. The overall workflow is summarised in Figure 3.8. Instead of taking a list of keywords we created a lexicon, since visual perception is not only dominant in written accounts, but also exhibits the highest lexical variety [*San Roque et al.*, 2015; *Winter et al.*, 2018]. To do so, we used a combination of ScenicOrNot rankings and descriptions of these photographs in the Geograph project. If a photograph was ranked higher than 7.62 (mean value of all ScenicOrNot rankings plus two standard deviations) on average, we added its description to our 'scenic corpus'. A small number of negative descriptions is typical in written texts [*Dodds et al.*, 2014]; therefore, we added descriptions to the 'unattractive corpus' if photographs were rated on average less than 2.82, which is a mean minus one instead of two standard deviations (Figure 3.7). This resulted in 4847 entries in the 'scenic corpus' and 26029 entries (instead of 1427 for two standard deviation) in the 'unattractive corpus' extracted from Geograph UK. From these corpora, we extracted all dependencies labelled as 'adjectival modifiers', using a dependency parser[9]. For example, from the phrase, 'stunning panoramic views', we extracted two pairs: 'stunning views' and 'panoramic views'. We controlled for the statistical significance of these pairs, comparing them to descriptions rated between 2.82 and 7.62 (Chi-test, $p < 0.005$, Figure 3.7), which resulted in final lexicons consisting of 184 scenic and 214 unattractive statistically significant pairs.

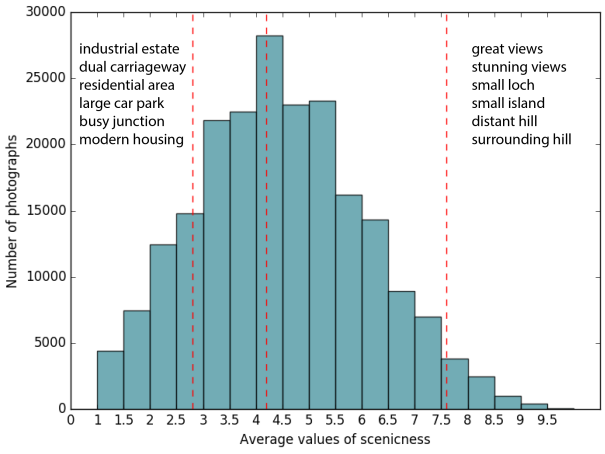


Figure 3.7: Counts of photographs rated in the ScenicOrNot project, red dashed lines correspond to mean value, mean value minus one standard deviation and mean value plus two standard deviations. Figure from Koblet and Purves (2020) – Publication 4.

9 https://spacy.io/usage/linguistic-features

The pairs of the unattractive lexicon are mostly related to urban landscapes (types of roads, car parks, buildings), and most of the pairs of the scenic lexicon are either related to abstract concepts (e.g., 'great views') or to specific terms describing landscape features and their properties (e.g., 'small island'). The extraction of all syntactic pairs labelled as 'adjectival modifiers' from our corpus of first-person landscape perception in the Lake District and their comparison with the pairs in the lexicon resulted in 28179 scenic and 266 unattractive descriptions, highlighting the above average scenicness of the Lake District.

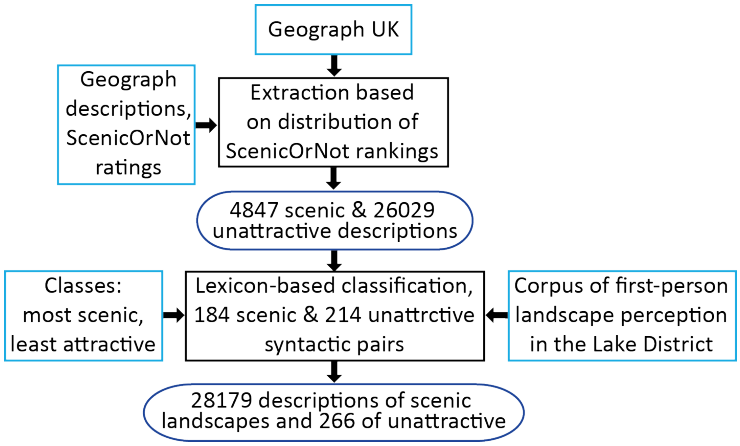


Figure 3.8: Overall workflow used to extract and classify descriptions of visual perception.

3.5.2 *Aural perception*

For aural perception we used classification based on sound emitters, as suggested in the field of ecoacoustics: anthrophony, biophony and geophony (Section 2.1.4). To these classes we added an additional category for the absence of sounds, an important aspect of landscape revealed by MacFarlane et al. (2004). The overall workflow is summarised in Figure 3.9. To extract a sub-corpus of aural descriptions, we assembled several lists of keywords that include verbs of sound emission, verbs of sounds made by animals and verbs of sound existence as listed in [*Levin*, 1993], their synonyms from WordNet [*Fellbaum*, 1998] and a list of adjectives related to sounds[10]. This sums to a total of 196 words. We performed the pre-processing steps as described in Section 2.3.2, with the most important steps being the reduction of the words to their basic forms in the process of lemmatisation and identification of their part of speech. Since words describing aural perception are highly polysemous, we used WordNet hypernyms and sentence context as implemented in the Lesk algorithm to disambiguate them (Section 2.3.3) [*Manning and Schutze*, 1999]. Since we aim for high precision we have iteratively extracted subsets and manually identified

10 https://www.sightwordsgame.com/parts-of-speech/adjectives/sound/

cases that commonly resulted in false positives. We then refined the rules eliminating such cases; for example, we did not use the keyword 'echo' as a verb, but kept it if it was a noun, since the majority of the descriptions with it as a verb were metaphorical, such as in 'echoes the style of Victorian buildings'. These heuristics resulted in the extraction of 8784 descriptions from the Geograph UK corpus with a precision of 0.75. Further, we annotated these descriptions with Cohen's Kappa inter-annotator agreement of 0.88 [*Landis and Koch*, 1977]. The classification step was then performed by using a random forest classifier tested and trained on the random halves of the 8784 descriptions with the following features: the 500 most frequent words, a list of British birds and mammals and a list of natural features and their qualities. The precision of this method is 0.81 and recall is 0.70. Our final results demonstrated that descriptions of aural perception are dominated by the absence of sound (5146), followed by anthrophony (2275), biophony (832) and geophony (386). These results and our micro-analysis during the annotation process revealed the complexity of the concepts of quietness, peacefulness and tranquillity, as described through natural language. To explore this phenomenon further, we decided to move from the class 'absence of sound' and extract descriptions referencing to a collective notion of tranquillity in the Lake District using the historical corpus CLDW and Geograph Lake District (Table 3.4).

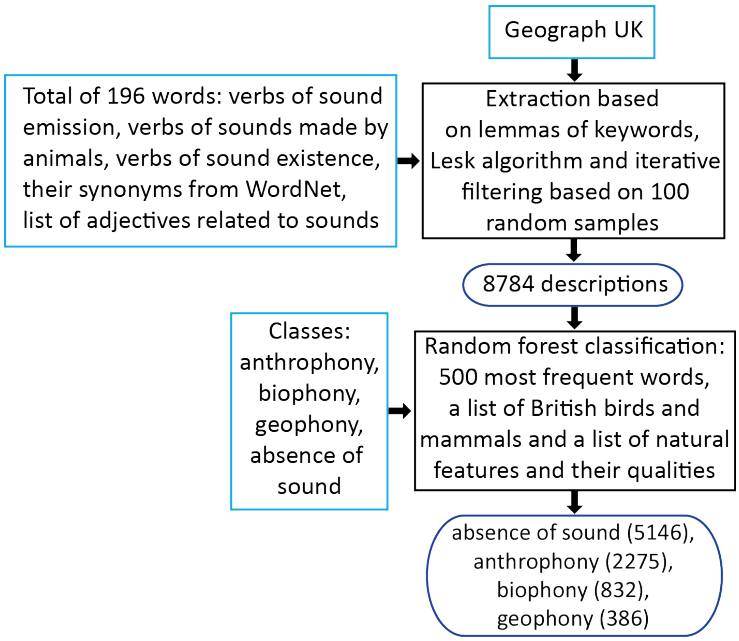


Figure 3.9: Overall workflow used to extract and classify descriptions of aural perception.

3.5.3 *Tranquillity*

To extract the intertwined references to tranquillity and the absence of sounds, we used a list of terms taken from the Historical Thesaurus of English[11] in the categories 'inaudibility', 'faintness/weakness' and 'quietness/ tranquillity', and filtered out terms which were only in use before 1750, resulting in a total of 66 terms (Figure 3.10). Since most of the words were adjectives, we performed a disambiguation step based on part of speech tagging (Section 2.3.3).

Classifications commonly used for tranquillity (continuous values from the least to the most tranquil landscapes) are not suitable when working with natural language. Therefore, keeping in mind studies demonstrating interplay between visual and aural perception [*Southworth*, 1969; *Carles et al.*, 1999; *Pheasant et al.*, 2008], we performed an iterative process of macro- and micro-analysis. Starting from our keywords, we explored co-occurrences between them, and other words found in the historical corpus CLDW and in the Geograph Lake District. These co-occurrences and our micro-analysis of the texts revealed that generic place descriptions (e.g., 'spot', 'scene') are often used to characterise both visual and aural perception as suggested by literature. However, anthropogenic objects (e.g., 'motorway', 'road') and references to time (e.g., 'morning') were clear indicators of a contrast between peacefulness of a certain location and anthropogenic intrusions nearby or between an early hour quietness and its later disturbance. Other terms, including 'calm' and 'peace' often co-occurred with weather and water-related words, implying not only an absence of sounds, but also of movement. Based on these explorations, we introduced the following classification of tranquillity:

- **Combination of visual and aural perception:** Descriptions, where visual attributes of the scene are as important as aural, and where absence of sounds is implicit (e.g., 'A remembrance service is held here every year and I can't think of a more beautiful and peaceful place to reflect.'[12])
- **Contrasting sounds:** Descriptions reflecting ephemerality of tranquillity by comparing it to other less tranquil locations, different time of day or mentioning sounds which add or detract from overall tranquillity (e.g., 'A moment of peace at Ashness Bridge – rare moments indeed!'[13])
- **No-movement**: Explicit mention of a lack of movement with implied silence and tranquillity (e.g., 'Below to the west Buttermere appeared mirror calm, the blue of the sky reflected deeply in its chill waters.'[14])

11 https://ht.ac.uk/
12 http://juliahedges.blogspot.com/2018/06/a-walk-up-mighty-great-gable.html
13 http://www.lakedistrict-walks.co.uk/2016/September/10.09.2016_Bleaberry_Fell.html
14 http://www.david-forster.com/section278539_221484.html

- **Total silence and tranquil sounds:** Either descriptions of tranquil sounds without contrast or explicit descriptions of complete silence (e.g., 'Further to the north Blencathra and Skiddaw put in an appearance in the evening sun, and we stopped to listen to the silence – not a sound – very peaceful and relaxing.'[15])

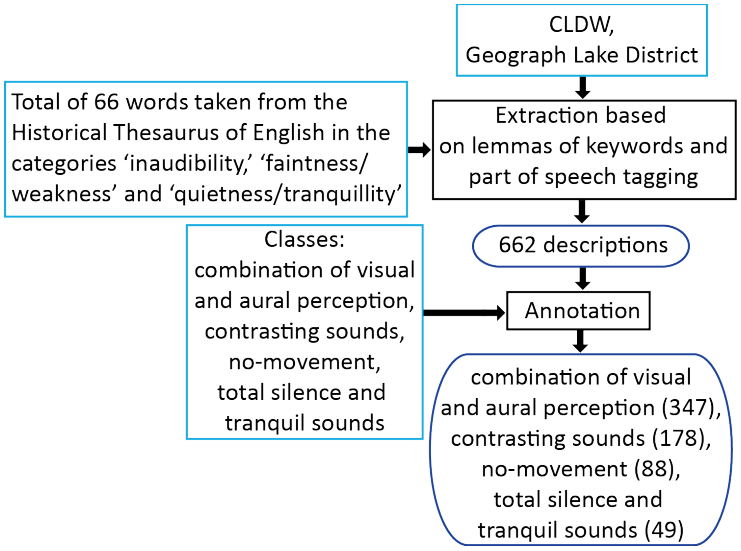


Figure 3.10: Overall workflow used to extract and classify descriptions of perceived tranquillity.

The annotation of these descriptions, especially in the historical corpus, was challenging even for human annotators with Cohen's Kappa inter-annotator agreement being 0.88 for the corpus Geograph Lake District and 0.62 for the CLDW [*Landis and Koch*, 1977]. Therefore, for the historical corpus the annotators worked together to come to a consensus for every description. In total, from both corpora we extracted 347 references to the combination of visual and aural perception, 178 descriptions of contrasting sounds, 88 of no-movement and 49 of total silence and tranquil sounds.

For our further experiment with the corpus of first-person landscape perception in the Lake District, we added together descriptions of sound experiences extracted from Geograph UK, Geograph Lake District and CLDW, which led to a total of 9446 descriptions used as training data. Then, we applied the same workflows as described in this section and in Section 3.5.2 and, first, extracted 1711 descriptions which were gradually reduced to 1480, since we had to adjust the lists of false positives to the new corpus (e.g., Quiet Garden is a name of a garden in Rydal Hall, which had to be filtered out unless other words referring to sound experiences were present in the sentence). These 1480 Lake District descriptions were classified as perceived tranquillity (886), geophony (278), anthrophony (174) and biophony (142) (Figure 3.11).

15 http://www.ramblingpete.walkingplaces.co.uk/day/lakes/martindale.htm

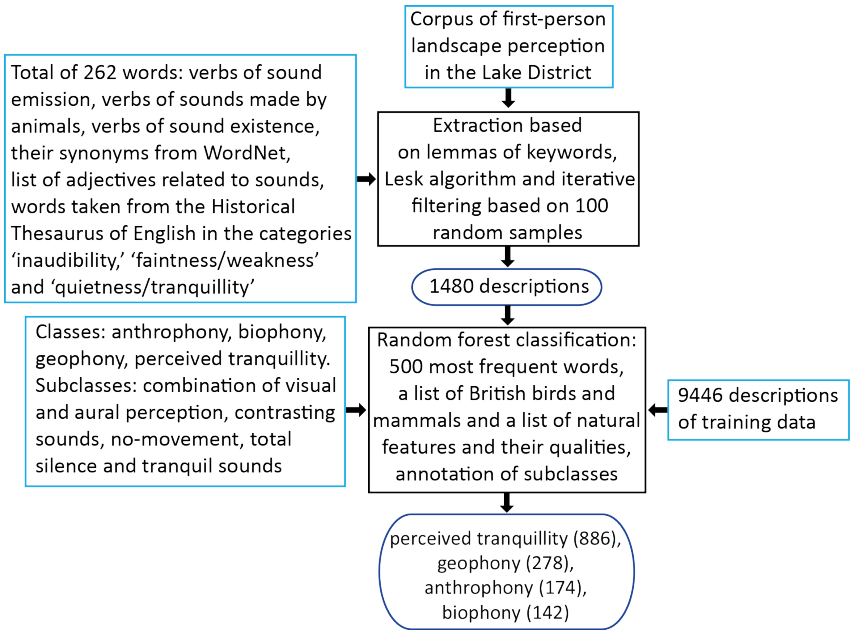


Figure 3.11: Final workflow used to extract and classify descriptions of aural perception and tranquillity.

3.5.4 *Olfactory perception*

To create a list of keywords for olfactory perception, we combined verbs of smell emission listed in [*Levin*, 1993] with WordNet lists indicated as 'olfactory property', 'malodour', 'acridity', 'aroma', and 'scent' [*Fellbaum*, 1998] and adjectives with 'olfactory' as the dominant modality [*Lynott and Connell*, 2009]. This resulted in a total of 29 words. Similarly to aural perception, we grouped olfactory perception based on smell emitters. The extraction step was performed only on the new corpus of first-person landscape perception in the Lake District, in which 78 olfactory descriptions were found. We annotated these as smells emitted by plants, by animals or as smells of an anthropogenic nature (Figure 3.12).

3.5.5 *Assigning texts to space*

In the Geograph corpus, descriptions are explicitly linked to coordinates. For the CLDW and the corpus of first-person landscape perception in the Lake District, we assign descriptions to space using two levels of granularity: areas of distinctive character for visual perception, and individual named landscape features (e.g., summit, valley) for descriptions of tranquillity, aural and olfactory perception.

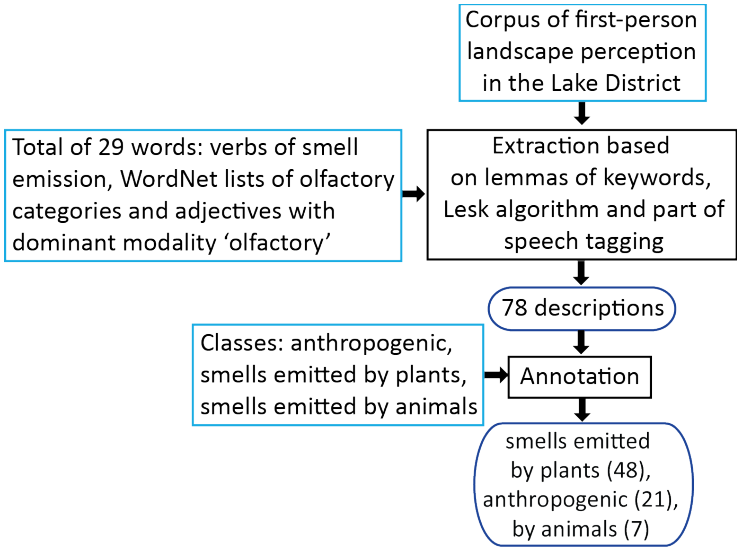


Figure 3.12: Overall workflow used to extract and classify descriptions of olfactory perception.

To assign textual documents to areas of distinctive character, we use the following heuristics:

- For the initial setting we count the frequency of each toponym in each text as demonstrated in Figure 3.13. Then, in the step 1, we apply density-based clustering [c.f. *Moncla et al.*, 2014]. Clustering allows us not only to disambiguate toponyms, but also to remove outliers such as distant peaks, since these do not typically form a cluster (e.g., Scafell in Figure 3.13, or distant locations used to describe a generic landscape element (e.g., 'the road from Keswick to Kendal' in Figure 3.13).
- Step 2: if all toponyms mentioned in the same text are within the same area, we assign this text to this area. If there is more than one area, we create three classes of toponym frequency based on Jenks natural breaks data clustering, and we further work only with the most frequent class (red and blue clusters in Figure 3.13). The assumption is that salient locations, which are only seen, but not visited, are not mentioned as often as those visited.
- Then, in the step 3, we take the most frequent toponym as our starting point to check if the rest of the toponyms from the most frequent class are in the neighbouring areas (red toponym with frequency 6 in Figure 3.13). This step is necessary, since travellers often walk on ridges, which potentially coincide with the borders of the areas of distinctive character. In such cases we assign the text to both areas (areas A and B in Figure 3.13). In case there is no unique most frequent toponym, we take as the starting area the one which also has the highest number of toponyms from the most frequent class.

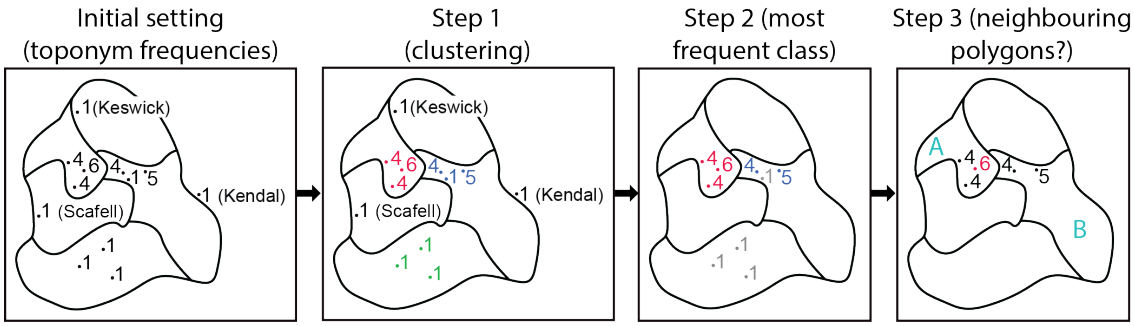


Figure 3.13: The workflow to assign documents to areas of distinctive character. Areas A and B are the final document scope, names in brackets refer to examples in the text. Figure from Koblet and Purves (2020) – Publication 4.

For individual landscape features there are three main cases:

- If there is a mention of one toponym in a sentence, we check for referent ambiguity, e.g., how many entries does this toponym have in the gazetteer? If there is only one entry, we simply assign coordinates of the gazetteer entry. If there is more than one entry, we extract the closest toponyms in text (before and after the initial one) and disambiguate the initial toponym based on the distance proximity [*Leidner et al.*, 2003].
- In the case of more than one toponym in a sentence, we select the one, which appears in the URL, since it is often the main goal of the journey. If there is no match, we select the toponym that has fewer counts in the gazetteer or randomly in the case of a tie.
- If there are no toponyms in the sentence, we repeat the process at the paragraph-level.

Having completed this process, we have a list of syntactic pairs from our lexicon (e.g., 'great view') associated with areas of distinctive character and a list of descriptions referring to tranquillity, aural and olfactory perception associated with individual landscape elements. To reduce the effects of user-generated biases, we retained only one description if there were several with the same class, location and author.

3.6 LANDSCAPE CHARACTERISATION

In the following, we characterise our study area (Great Britain) and case study area (Lake District) in respect to perceived landscape properties. We will move from the coarse granularity of Great Britain through the Lake District as a whole and areas of distinctive character to individual landscape elements.

3.6.1 *Great Britain*

Some patterns become apparent already on the granularity level of the whole region, which we were able to extract from Geograph UK. Visual perception varies strongly within Great Britain, with clusters of less scenic landscapes around urban centres, such as London, and the most scenic locations in Scotland to the north-west of Edinburgh and Glasgow (Figure 3.5). These results for Great Britain can be modelled based on the votes of the ScenicOrNot project alone, and, indeed, such research exists [e.g., *Seresinhe et al.*, 2015]. However, textual descriptions reveal other insights into the ways landscapes are perceived.

We grouped nouns in word clouds (Figure 3.14), depending on average scenicness ratings of photographs and their associated descriptions. The nouns in the lowest scenicness category exhibit a clear trend towards developed areas, with the words 'motorway', 'store' and 'factory' being frequent in this class. The nouns in the highest-rated class are less common and are related to natural processes (e.g., 'avalanche'), wildlife (e.g., 'otter') and Gaelic words (e.g., 'mhonaidh'). Nouns in these two classes come from a small number of photographs, 14072 (ca. 6%) in the lowest and 3134 (ca. 1%) in the highest class; therefore, one must be cautious with the interpretation of this data.

(a) Scenicness between 1 and 3, 14072 images, 2137 users

(b) Scenicness between 3 and 5, 155822 images, 4170 users

(c) Scenicness between 5 and 7, 79752 images, 3340 users

(d) Scenicness between 7 and 10, 3134 images, 851 users

Figure 3.14: Average scenicness for 150 most frequent nouns extracted from image descriptions, font size indicates relative frequency within scenicness range. Underlined words are discussed in the text. Figure adapted from Chesnokova et al. (2017) – Publication 1.

The majority of the nouns belong to another two classes, rated between 3 and 5 and between 5 and 7. This is not surprising, since each photograph was rated at least three times, and many nouns appear in multiple descriptions. These two classes do not exhibit such a clear distinction as those with the least and the highest scenicness, but we still see a difference of nouns related to rural landscapes (e.g., 'hamlet', 'farm', 'wood') rated as less scenic than nouns related to perceived natural scenes (e.g., 'moorland', 'mountain', 'ridge').

Since we extracted descriptions of aural perception for the whole region of Great Britain and Ireland, we can give some characterisation of this region from this perspective as well. The majority of the sound-related descriptions refer to perceived absence of sound (59%), followed by anthrophony (26%), biophony (10%) and geophony (5%). Figure 3.15 demonstrates the spatial distribution of references to perceived absence of sound through aggregated counts of descriptions for 20 square kilometre grid cells. Individual descriptions give an additional level of understanding through micro-analysis of what is considered important for the authors describing this region.

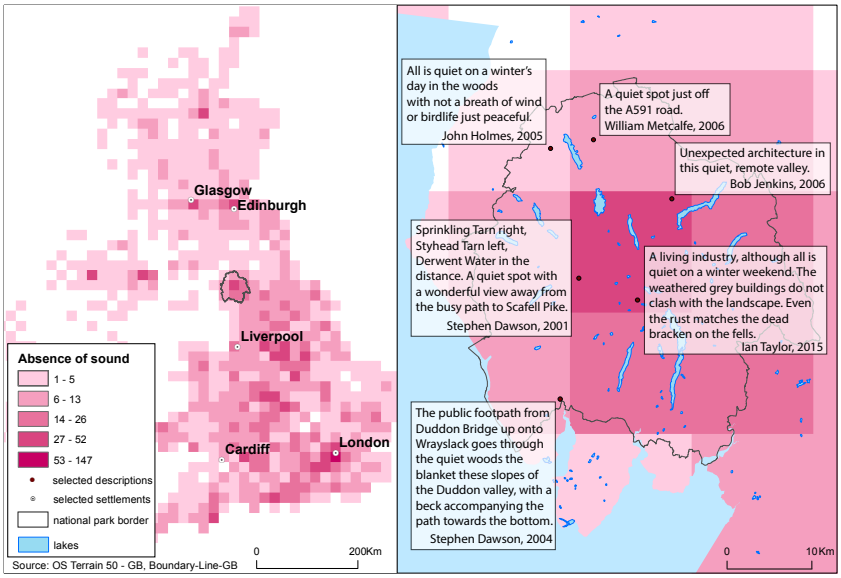


Figure 3.15: Aggregated number of descriptions related to absence of sound (macro-analysis) with selected descriptions for micro-analysis.

In Section 2.1.4 we described the complexity of perception of natural and human-generated sounds. To test which types of sounds and their emitters are perceived more positively or more negatively, we performed sentiment analysis of our descriptions using a procedure outlined by Iyyer (2015), Section 2.3.5. Since the output of this approach is average sentiment value per sentence on a continuous scale, we generated three classes of negative, positive and neutral sentiments by taking half a standard deviation less than the mean, half a standard deviation more than the mean, and values in between, respectively. Figure

3.16 shows the resulting distribution, with geophony and biophony, contrary to our expectation, showing strong association with negative sentiments. Absence of sound, however, is dominated by neutral and positive sentiments. As in the case of visual perception, we created word clouds to explore what kind of concepts have negative or positive connotations.

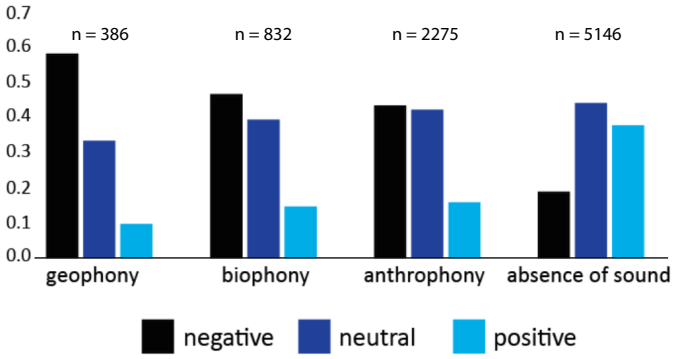
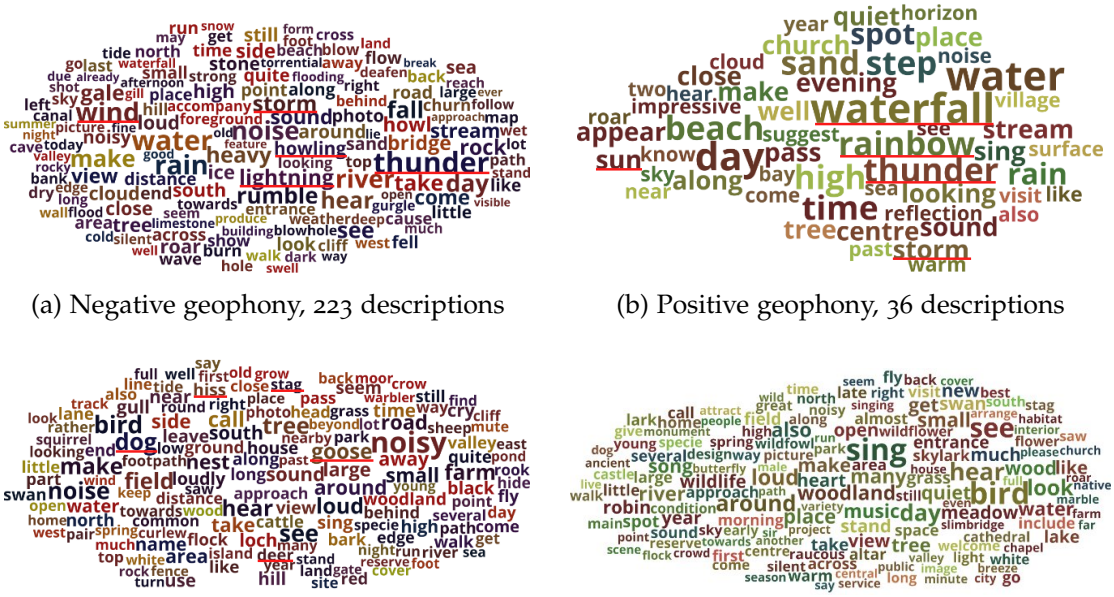


Figure 3.16: Proportion of descriptions according to sentiment values. Figure adapted from Chesnokova and Purves (2018) – Publication 2.

In geophony, words related to weather are not uniquely negative or positive; the words 'storm' and 'thunder' appear in both classes (Figure 3.17). However, in the negative class they co-occur with 'lightning', 'wind' and 'howling', whereas in the positive class – with 'rainbow', 'sun' and 'waterfall'. This difference suggests that, in the positive class, mostly the events after the storm are described and that 'thunder' may be a reference to the sublime sounds of waterfalls. Negative biophony contains words related to farm animals, e.g., 'goose', 'dog', 'hiss', and words related to roaring stags, whose sounds may be perceived as scary, e.g., 'stag', 'deer'. Positive biophony is quite different; we observe more words related to singing birds and wildlife overall. To compare overall patterns of Great Britain to individual regions, we analysed descriptions within the boundaries of the UK's 15 national parks [*National Parks UK*, 2019]. These revealed that the class absence of sound in national parks exhibits clear negative sentiments towards human activity and traffic (e.g., 'railway', 'traffic', 'warehouse') and positive sentiments towards natural land forms (e.g., 'summit', 'beach'), generic locations (e.g., 'scene') and adjectives related to the concept of tranquillity with ca. 67% of the occurrences of the adjective 'isolated' within the boundaries of national parks in comparison to Great Britain as a whole, ca. 40% of the adjective 'remote' and ca. 20% of the adjective 'tranquil'.



(a) Negative geophony, 223 descriptions

(b) Positive geophony, 36 descriptions

(c) Negative biophony, 386 descriptions

(d) Positive biophony, 119 descriptions

Figure 3.17: 150 most frequent tokens occurring more than three times for negative and positive descriptions of geophony and biophony. Underlined words are discussed in the text. Figure adapted from Chesnokova and Purves (2018) – Publication 2.

3.6.2 *Lake District*

At the granularity of the Lake District, using our lexicon and the corpus of first-person landscape perception in the Lake District, we extracted 28179 descriptions related to scenic landscapes and only 266 describing unattractive locations. This difference can be explained by the above average attractiveness of landscapes in the Lake District, since the lexicon is based on the descriptions of the whole island. Figure 3.18 demonstrates that modelled scenicness in the Lake District almost lacks the lowest class values, and the following classes of scenicness comprise ca. 47%, ca. 30% and ca. 8% with no grid cells classified as the highest class (Section 3.4).

We summarise the ten most frequent syntactic pairs in both classes in Table 3.6. Scenic pairs describe varieties of 'views'[16] and experiential properties of landscapes, such as 'ascent'[17] and 'descent'[18]. We see that uncommon, but still

16 'It was well worth the effort for the great views of Long Sleddale in both directions.', source: http://www.mypennines.co.uk/lake-district/walks/290106.html#sthash.rRTPr4JF.dpbs

17 'Decided to go via steep pathless ascent through heather whilst enduring horizontal rain, as a surprise challenge from the walk leader.', source: https://www.hill-bagging.co.uk/mountaindetails.php?qu=B&rf=3741

18 'Heading south from Keswick, there's a short sharp final climb up towards the summit and a very steep, rocky descent.', source: http://brianwalking.blogspot.com/2011/03/lake-district-walla-crag.html

present, unattractive syntactic pairs relate mostly to transport (e.g., 'parked cars'[19]) and previous industry in the region (e.g., 'old machinery'[20]).

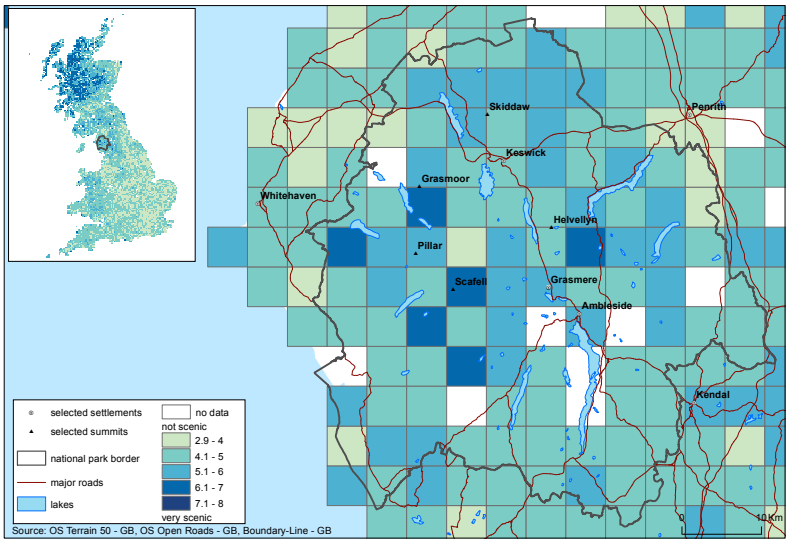


Figure 3.18: Predicted scenicness of the Lake District in comparison with Great Britain.

Table 3.6: Ten most frequent syntactic pairs from scenic and unattractive lexicons.

#	Scenic pairs	Count	Unattractive pairs	Count
1	great views	1012	large (car) park	39
2	highest point	881	parked cars	26
3	good views	707	dual carriageway	18
4	steep descent	528	old works	12
5	steep ascent	370	old machinery	10
6	good path	353	adjacent (car) park	6
7	good view	315	local shops	6
8	stunning views	314	static caravans	6
9	great view	306	much traffic	6
10	lower slopes	296	second bridge	5

For aural perception we extracted 1480 descriptions from the corpus of first-person landscape perception in the Lake District, with the majority of them

19 'After leaving there it was all downhill to Aira Force - with its hundreds of double parked cars and heaving summer hoards.', source: https://oldrunningfox.blogspot.com/2018/08/a-lakeland-jaunt.html

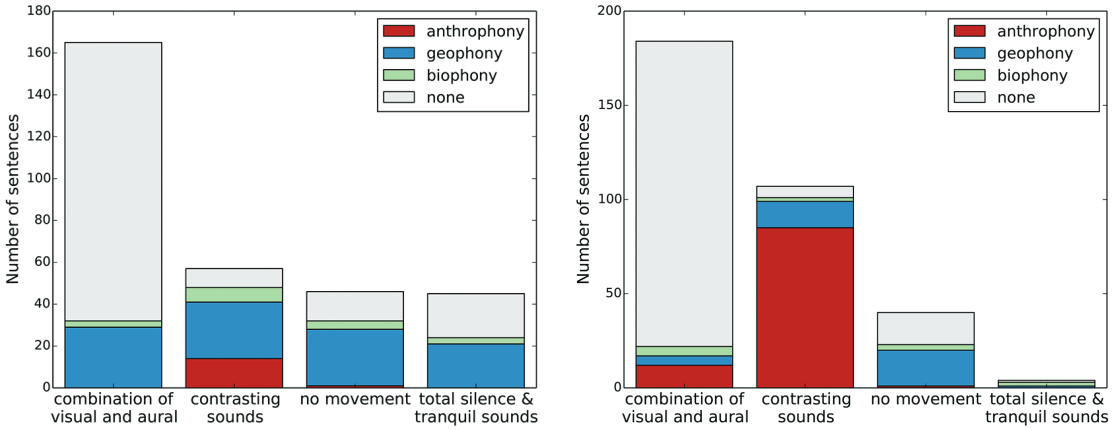
20 'Dotted all around are spoil heaps, rusting iron cables lie along the path, bits of old machinery lay abandoned on the mountainside, and a metal tower from an aerial tramway lays toppled on its side.', source: https://notesfromcamelidcountry.net/category/coniston/

referring to perceived tranquillity and the general absence of sound (ca. 60%) (Table 3.7), similar to the results we obtained for Great Britain as a whole. Within this class, the most common are combinations of visual and aural perception, with only half as many contrasting sounds; but these contrasting sounds particularly reveal the importance of using natural language in the explorations of human landscape perception. To explore them further, we extracted references to contrasting sounds from the historical corpus CLDW and from Geograph Lake District.

Table 3.7: Summary of extracted descriptions of aural perception and tranquillity from the corpus of first-person landscape perception in the Lake District.

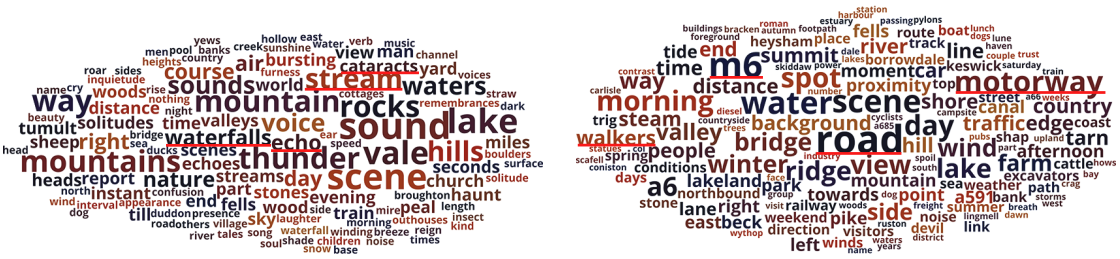
Type of sound experience	Count
Combination of visual and aural perception	485
Contrasting sounds	275
No-movement	66
Tranquil sounds and total silence	60
Total perceived tranquillity	886
Anthrophony	174
Biophony	142
Geophony	278
Total assigned to emitter	594

Figure 3.19 breaks down the tranquillity class further, showing the proportion of individual types of sound emitter per class. Striking here is the change over time of the proportion of anthrophony in the contrasting sounds. This difference reminds us of William Wordsworth and his tradition of describing Lake District landscapes by emphasising the tranquil sounds of flowing water (geophony) and overlooking sounds of industrial activity. In Figure 3.20, we see that, in the CLDW corpus, references to 'stream', 'waterfalls' and 'cataracts' are frequent. We also see the word 'echo', which gave the listeners exactly this contrasting opportunity of perceiving sounds, since echoes (especially those of canons) alternate with moments of silence, making the silence noticeable [*Taylor*, 2018]. In Geograph, however, tranquillity is valued by contrasting it to anthropogenic intrusions, such as traffic noise (e.g., 'road', 'motorway' and 'm6') and other visitors (e.g., 'people', 'walkers'), references to which are seen in Figure 3.20.



(a) CLDW (b) Geograph

Figure 3.19: Number of sentences per tranquillity class grouped by sound emitters. Figure from Chesnokova et al. (2019) – Publication 3.



(a) CLDW (b) Geograph

Figure 3.20: 150 most frequent nouns extracted from the class of contrasting sounds. Underlined words are discussed in the text. Figure adapted from Chesnokova et al. (2019) – Publication 3.

Comparing spatial distribution of extracted tranquillity with a map of independently modelled relative tranquillity created by MacFarlane et al. (2004) revealed that classes of no-movement and combination of aural and visual perception spatially coincide with this model that uses proximity to potential noise emitters (e.g., motorways) as a proxy for lack of tranquillity (Figure 3.21). The class of contrasting sounds is associated with lower values of relative tranquillity (Kruskal-Wallis test, $p < 0.01$), since such pockets of tranquillity are typically close to motorways and, thus, contrasted to anthropogenic intrusions[21].

21 'A tranquil scene viewed from the not-quite-so-tranquil A6 on the edge of Milnthorpe', source: https://www.geograph.org.uk/photo/2905720

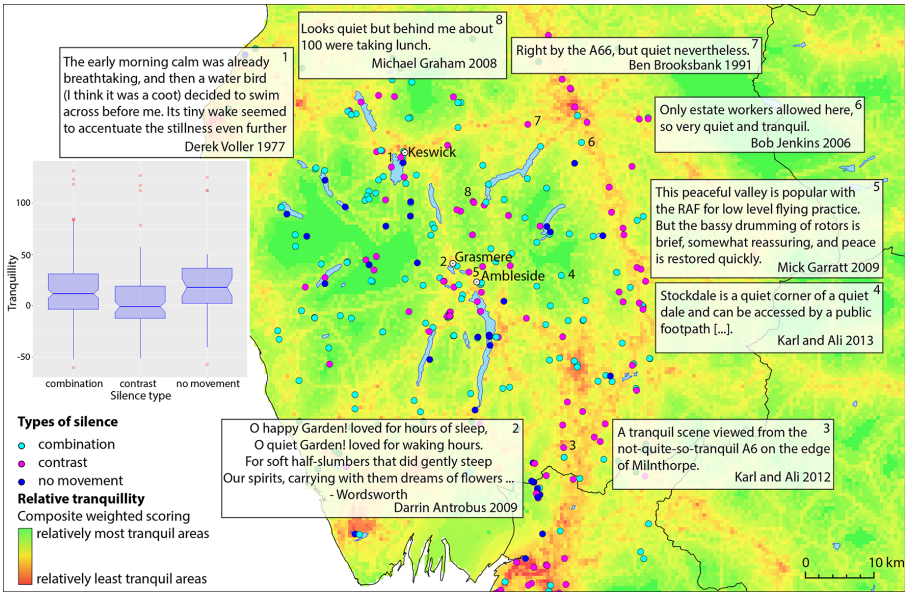


Figure 3.21: Comparison of the map of relative tranquillity [*MacFarlane et al.*, 2004] and types of tranquillity extracted from Geograph Lake District. National Tranquillity mapping data 2007 developed for the Campaign to protect rural England and Natural England by Northumbria University. OS Licence number 100018881. Figure from Chesnokova et al. (2019) – Publication 3.

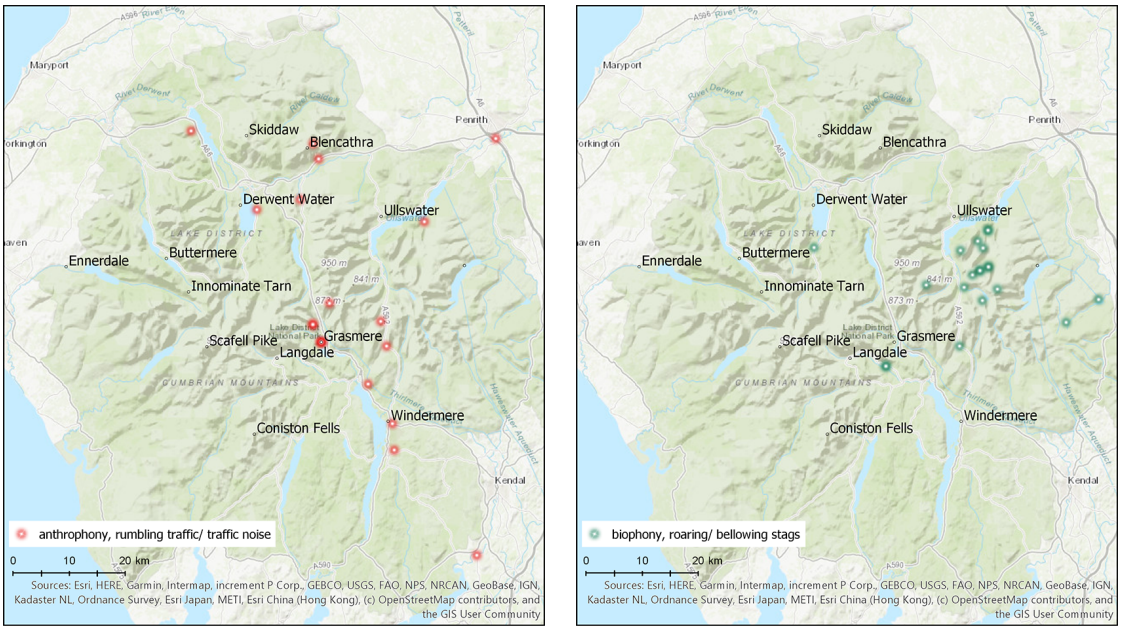
Coming back to the corpus of first-person landscape perception in the Lake District and looking at these intrusions spatially (Figure 3.22), we see that, for example, anthrophony, describing rumbling traffic and traffic noise is experienced not only in the valleys, but also on summits close to main transportation lines. The map of biophony demonstrates another relationship between aural perception and the Lake District; the cluster south of Ullswater is an old red stag area in England, which emerges also through such ephemeral perception as aural[22].

Of 78 extracted descriptions of olfactory perception, 48 describe smells emitted by plants (e.g., heather, juniper), 21 by anthropogenic sources (e.g., food, smoke) and 7 by animals. Spatially, smells are distributed over the whole area of the Lake District.

3.6.3 *Areas of distinctive character*

There are 71 areas of distinctive character within the borders of the Lake District National Park, and for 54 of them we were able to collect more than 10 texts using our corpus of the first-person perception in the Lake District. The areas in the south of the region are not covered by our corpus, since we used

22 'This is the oldest red stag area in England and the "Rut" had began as roaring stags could be heard all around.', source: http://frasermackay.blogspot.com/



(a) Anthrophony, rumbling traffic and traffic noise, 19 descriptions

(b) Biophony, bellowing/ roaring stags, 20 descriptions

Figure 3.22: Spatial distribution of selected perceived sounds in the Lake District. Figure from Koblet and Purves (2020) – Publication 4.

Wainwrights as search terms, and Wainwright described the area to the south in another book, *The Outlying Fells of Lakeland*. To illustrate how we can use references to landscape perception in our corpus, we compared three distinctive character areas with each other:

- 'Scafell Massif' (ca. 101 square km) is an area in the central part of the Lake District, containing its and England's highest summit – Scafell Pike.
- 'Skiddaw and Blencathra' (ca. 113 square km) is an area in the North-East of the Lake District, containing Skiddaw – one of the four English summits above 3000 ft.
- 'Upper Windermere' (ca. 27 square km) is an area containing a popular touristic settlement, Ambleside, the National Park visitor centre, and the steep Kirkstone Pass.

For visual perception, we selected 34 areas, setting the minimum number of texts per square km to two texts. For these 34 areas, we created word clouds based on the 50 syntactic pairs having the highest normalised score, for which we used spatial tf-idf [*Rattenbury and Naaman*, 2009]. This allowed us to reveal syntactic pairs that are frequent in one area but not in others and down-weight the pairs that are frequent in all areas. We found the following similarities and differences of these areas:

- The areas of 'Scafell Massif' and 'Skiddaw and Blencathra' contain only syntactic pairs from the scenic lexicon, whereas the area of 'Upper Win-

dermere' is also described by pairs related to developed areas and transportation axes[23].

- Both 'Scafell Massif' and 'Skiddaw and Blencathra' have 'highest point' as the highest ranked pair; however, only 'Scafell Massif' has prominent both 'highest mountain'[24] and 'highest peak'. The relatively high elevation of these areas is also revealed by the pair 'deep snow'[25].
- Less highly ranked but unique to individual areas are 'tussocky grass'[26] and 'covered slopes'[27] for 'Skiddaw and Blencathra' and 'small hill'[28] for 'Upper Windermere'.
- Other pairs 'small cairn' and 'true summit' reveal the hill-bagging nature of the region and the importance of Wainwright's books[29].

Out of 1426 texts describing 'Scafell Massif' we were able to extract 139 descriptions related to aural perception (ca. 10%) and similarly for 'Upper Windermere' (45 out of 338 descriptions, ca. 13%); however, for 'Skiddaw and Blencathra', this proportion is much higher – 122 out of 581 descriptions (ca. 21%). These three areas have the majority of descriptions classified as tranquillity (Table 3.8). However, differences within tranquillity classes exist; 'Upper Windermere' has a similar number of descriptions in the class of combination of aural and visual perception and in the class of contrasting sounds, whereas 'Skiddaw and Blencathra' have more than twice as many descriptions classified as combination, rather than contrast. Tranquil sounds and total silence also have different proportions in these areas; there are 12 descriptions in 'Skiddaw and Blencathra' and only 4 in 'Scafell Massif'.

In contrast to the absence of anthropogenic syntactic pairs in visual perception for the 'Scafell Massif' and 'Skiddaw and Blencathra' areas, we see that anthropogenic sounds are present in these areas, especially in 'Scafell Massif', where low flying jets are mentioned 10 times. References to traffic were present already

23 'There were several paths we could have taken, but we chose the Underbarrow road because it has the advantage of a bridge over the busy A591 dual carriageway.', source: https://beatingthebounds.wordpress.com/2016/04/12/silverdale-to-keswick-ii-to-ambleside/

24 'On a clear day, England's highest mountain, Scafell Pike, affords spectacular views across England's deepest lake and the surrounding Lakeland fells.', source: https://lagrenouilleanglaise.wordpress.com/tag/scafell-pike/

25 'We followed the well worn path towards High Raise (2500ft/762m) which was covered in deep snow.', source: https://gappet.wordpress.com/category/2014/trail-100/page/4/

26 'Above White Gill the slope eases on the approach to the summit plateau, a dreary expanse of tussocky grass and sphagnum moss', source: http://www.wainwrightroutes.co.uk/mungrisdalecommon_r1.htm

27 'Heather-covered slopes of Carlside behind, Skiddaw Little man behind that.', source: http://josweeney.net/ullock-pike-to-dodd-in-snow/

28 'This small hill all 1,588ft of her introduces many people to the fells, a changer of lives once climbed never forgotten', source: http://www.one-foot.com/Wansfell%20Pike%20via%20Jenkin%20Crag%20and%20Troutbeck%202010.html

29 'From there we followed the crest of the ridge for 1 mile to get to the true wainwright summit of Baystones Seeing Red Screes and the far eastern fells in all their glory', source: https://www.hill-bagging.co.uk/mountaindetails.php?qu=S&rf=2607

in the references to visual perception of 'Upper Windermere' and appear again with 4 mentions of traffic noise and an overall high proportion of anthrophony (20%). Geophony is present, especially in high-lying areas, with references to howling wind and crunchy snow[30].

Table 3.8: Summary of extracted descriptions of aural perception and tranquillity in individual areas of distinctive character.

Type of sound experience	Scafell Massif	Skiddaw and Blenacathra	Upper Windermere
Combination (visual and aural)	41	48	13
Contrasting sounds	32	22	11
No-movement	5	0	1
Tranquil sounds and total silence	4	12	2
Total perceived tranquillity	82	82	27
Anthrophony	18	8	9
Biophony	11	9	3
Geophony	28	18	6
Total assigned to emitter	57	35	18

References to olfactory perception are relatively rare; however, they give additional characterisation of areas of distinctive character, with, for example, 4 of 9 mentions in the area of 'Skiddaw and Blencathra' referring to the smell of heather[31] on the 'heather-covered slopes'. By contrasting neighbouring areas to each other, we see that 'Upper Windermere' has only one reference to smells and this is of anthropogenic nature of a fish and chips shop, whereas neighbouring 'Grasmere and Derwent Water' has 7 descriptions, with 4 of them referring to scents of plants. 'Scafell Massif' has a concentration of smells on the steep Hardknott Pass, with 4 of 5 descriptions referring to the smell of burning brakes. However, this information reveals more about the individual landscape element than about the area as a whole, making it important to look at this level of granularity separately.

30 'It was only after leaving the steepness of Further Gill Sike could I start to really enjoy my walking again, here I am met by scattered snow along the ground which had no real depth to speak of, it just made that reassuringly crunching sound as I walked over it.', source: http://sharkeysdream.co.uk/PAGES/WALKS/20131226.html

31 'Descending Carl Side through heather in full bloom, the scent was just lovely.', source: https://www.wainwrightwalking.co.uk/ullock-pike-to-dodd/

3.6.4 *Individual landscape elements*

We applied the finest level of granularity in our analysis only to tranquillity, aural and olfactory perception extracted from the corpus of first-person landscape perception in the Lake District, omitting visual perception. We assume that such detailed analysis of vistas is less meaningful than analysis of sounds' and smells' emitters, though they too might vary depending on their static or dynamic nature and dependence on affordances (e.g., a road for traffic noise).

Looking at this level of granularity, important patterns emerge of the ways landscapes are experienced. Six out of ten most frequent mentions of landscape elements refer to lakes and tarns, with the majority of the descriptions classified as references to tranquillity, highlighting the importance of water as its constituent (Table 3.9). However, these locations also experience anthropogenic intrusions in the forms of traffic noise[32], people[33] and screaming jets[34]. One of the tarns – Angle Tarn – is situated at a higher elevation (479m) and we see similar references to howling wind as on the summit of Blencathra (868m). These two locations also have references to biophony in the form of roaring stags[35] and migrating geese[36]. Aira Force waterfall and Ashness Bridge are popular touristic destinations, both characterised by geophony, but of very different kinds. Aira Force is described using words referring to experiences of sublime, such as 'thundering' and 'deafening'[37] sounds similar to low flying jets as suggested by Fisher (1999), whereas sounds of the stream below Ashness Bridge evoke the perception of beautiful landscapes with de-stressing and wonderful sounds[38].

32 'Along the busy road between Grasmere and Ambleside, I permit myself the iPod ("Thick as a Brick") to cut out the traffic noise and to keep me trudging along even though my boots are pinching my toes.', source: http://www.zanthan.com/wordsintobytes/postcards/lake-district-silver-how/

33 '...the walk back to Buttermere village was further than it could have been it wasn't too bad at all, despite the numbers of people there whose limits of entertainment and exercise seemed to involve loudly throwing stones into the lake.', source: https://214wainwrights.wordpress.com/walk_list/walk16/

34 'The RAF have also been busy this week, the relative calm is broken every so often by a couple of jets screaming up the valley, sometimes they get very low...', source: https://geraldinebunn.wordpress.com/

35 'The head of Bannerdale, a stunning remote valley with the sound of Red Deer stags roaring all around', source: http://walksnwildlife.blogspot.com/2010/10/circuit-of-martindale-and-lots-of-red.html

36 'few things are more evocative of the phrase "Winter is Coming" than the sight of skeins of geese on the move', source: http://cheviots.blogspot.com/2015/10/bakestall.html

37 'Heavy rain in the days before our visit meant Aira Beck was in full flow and the noise was deafening!', source: https://these8boots.wordpress.com/

38 'We all agreed that it's the best sound for anyone looking to de-stress and for me nothing captures natures sound better.', source: https://upnoutside.wordpress.com/2012/01/21/ashness-bridge-walla-crag-keswick-saturday-21st-jan-week-4/

Table 3.9: Ten most frequent mentions of landscape elements in the descriptions of tranquillity and aural perception and their counts.

Landscape element	Count	Landscape element	Count
Grasmere (lake)	20	Angle Tarn	11
Buttermere (lake)	18	Ashness Bridge	10
Derwent Water	15	Ambleside town	10
Aira Force waterfall	13	Easedale Tarn	10
Blencathra fell	12	Haweswater	10

Olfactory perception also partly follows popular touristic locations; next to Aira Force, there is a smell of pine and fir[39], and there are two descriptions of the Gingerbread shop in Grasmere, with its 'intoxicating aroma'[40]. As was mentioned in the previous section, the highest number of olfactory references describe the smellscape of the steep Hardknott pass, with burning brakes and clutches[41]. These references describe the same landscape element – the Hardknott pass, but some references to smells are rather characteristic of the same type of landscape, such as cliffs or slopes. For example, a strong stench of dead sheep[42] is common next to steep cliffs, and the fragrance of heather is present on the slopes.

3.6.5 *Face validity*

To explore the practical utility of our approach, we built a group for face validity through an expert discussion [*Rykiel*, 1996]. With the help of the Lake District National Park Authority, we sent invitations to organisations locally (e.g., Friends of the Lake District) and England-wide (e.g., Natural England, Forestry Commission). However, only participants from local organisations had time to be part of the discussion (3 people). They were presented with explanations of sources and methods and with maps showing examples of our results. The maps contained the following information (Figure 3.23):

- Locations of initial search terms.

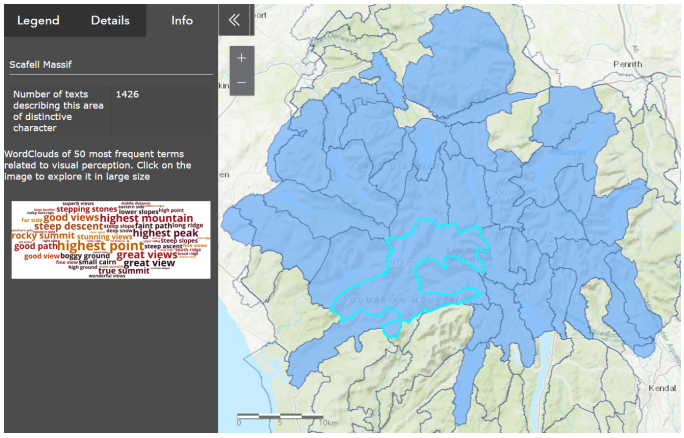
39 'Deep green pine and fir perfume the damp air with the heady scent of resin, conjuring memories of childhood Christmases.', source: https://amandaragaa.com/2018/03/29/aira-force/

40 'The first clue to the delicious treats inside this tiny place is the intoxicating aroma of ginger, cinnamon and sugar which is carried along the street on an inviting breeze.', source: https://amandaragaa.com/2015/10/23/grasmere-village-cumbria-wordsworth-country/

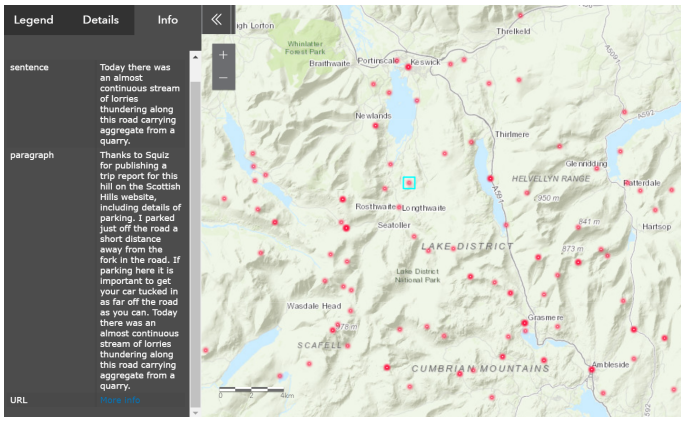
41 'Cars on 3 wheels coming around the steep hairpins, stinking of burning clutches and brake discs.', source: https://babtestingground.wordpress.com/2012/10/

42 'Unfortunately there was the stinking, rotting corpse of a sheep partly obstructing the path on the far side.', source: http://www.boydharris.co.uk/w_bh10/100512.htm

- Official geometries of areas of distinctive character with number of texts collected in this work, and word clouds representing visual perception for 34 of the areas, displayed by a click on each area.
- Point locations of classified aural and olfactory perceptions with the extracted sentences, paragraphs and the original URL available on a click.



(a) Visual perception



(b) Aural perception (anthrophony)

Figure 3.23: Interface of the maps presented to the expert group available under tinyurl.com/LakeDistrictPerception. The information on the left is displayed for area or point location selected by the user (highlighted in light blue). Figure from Koblet and Purves (2020) – Publication 4.

Participants were asked two groups of questions. The first was in the form of a SWOT analysis (Strength, Weaknesses, Opportunities and Threats), and the second was a set of more specific and detailed questions. The results of the SWOT analysis are summarised in Table 3.10. The overall feedback from the expert discussion was positive, and two major points were repeated: the potential of our methods for robust monitoring of changes and the concern about identification of groups of users whose opinions are covered by our descriptions. The value of our approach was explicitly stated not only for LCA, but also for other activities of the National Park authorities, such as access management, where

sentiments towards the problem before and after management activities would be very valuable. For example, the National Park has experienced an increase in 4x4 off-road vehicles near High Tilberthwaite farm, which were claimed to ruin the tranquillity and beauty of places once owned by Beatrix Potter and described by Alfred Wainwright as 'scenically one of the loveliest in Lakeland', leading to a threat of revoking the UNESCO World Heritage status [*Parveen*, 2018]. This year, a Traffic Regulation Order may be imposed on the two roads in question [*National Trust*, 2019b], making it interesting to see if perception of this location is going to change. Other examples include opinions over longer-term activities, such as forest plantations initiated by the Forestry Commission, and perception related to light pollution, where measurement results initiated by the Friends of the Lake District in the project 'Dark Skies' could be compared with people's perception[43] [*Korndorfer*, 2019].

For the detailed questions, the following topics have emerged:

- What makes a location unattractive for visitors is important and needs further detailed analysis.
- Word cloud visualisations could be used by communication teams, since they show what people are actually saying about the place.
- Looking at residents versus tourists would be valuable (even though there are also subgroups, e.g., people moving to the Lakes when they retire and people growing up there).

Overall, the experts' feedback was positive on both written first-person narratives as a data source and on our workflows of extraction, classification and linking descriptions to space. We discuss further some of their concerns in Section 4.1 and suggest new possibilities for monitoring in Section 4.4.

43 'From Troutbeck via bridleway and south ridge. Should have descended in darkness but light pollution, especially from Limefitt Holiday Park, ensured that it was nothing like proper night time.', source: http://www.hill-bagging.co.uk/googlemaps.php?qu=S&rf=2626

Table 3.10: Summary of the SWOT evaluation through expert discussion.

Question	Feedback
Strengths	The methods are repeatable and automatic, making the monitoring of e.g., every 5-year period possible. Methodology allows us to include other search terms (e.g., other than Wainwrights), which makes the method robust. Value for LCA was explicitly stated.
Weaknesses	Weaknesses of sources include potential bias to positive descriptions. Missing processing steps include identification of the date of the experience, since this information is very important for the monitoring and for the estimates of differences based on seasonality. Weaknesses in methodology include the following concerns: in the case of peace and tranquillity we don't know what the author's reference is (e.g., New Yorker tranquillity versus Highlander tranquillity). Descriptions by different user groups are mixed together, whereas looking at them separately would be valuable. People visiting fells or just looking at them should also be separated (visiting the town of Keswick and describing views of Skiddaw or being on the summit of Skiddaw should be two different types of descriptions).
Opportunities	The main opportunity is monitoring of landscape quality and how its improvement or decline affects perceptions. Monitoring of people's opinions towards access management actions and other planning decisions (e.g., before/after) is valuable. Another potential is identification of topics and species (e.g., alpine plants on Scafell), which are never perceived, but are very important from the ecological view. These identified topics can be used for further education of visitors. Other sources (e.g., Twitter, Facebook groups) will potentially show more negative and instantaneous opinions (e.g., traffic jams during the Bank Holidays).
Threats	Potential over/under-representation of certain groups of users (e.g., fell climbers), therefore, it has to be clearly communicated whose opinion is taken into account.

4

DISCUSSION

Our aim in this work was to demonstrate how information extracted from first-person narratives can enhance our understanding of human experiences of landscapes through a bottom-up process and thus complement current methods of landscape assessment. By adopting a variety of methods from GIScience and NLP, focusing on creation of a text corpus, extraction of relevant text snippets and their further classification and assignment to space, we were able to create a reproducible workflow and thus fulfil this aim. In this chapter, Sections 4.1-4.4 address the research questions identified in Section 2.4.

4.1 TEXTUAL CORPUS FOR FIRST-PERSON LANDSCAPE PERCEPTION (RQ1)

An important contribution of this work is a workflow that allows us to automatically collect spatially referenced and thematically relevant texts. We applied this workflow to create a corpus of ca. 7000 rich texts of almost 8 million words describing first-person perception in the Lake District, and spatially linked them to areas of distinctive character and individual landscape elements. This shows the transferability of our previous results based on descriptions collected in the Geograph project to other textual sources, which can be collected for any English-speaking country. We did not test it for any other languages, but if natural language processing tools are available in the language of interest, the workflow can be adapted. The size of our corpus is comparable to other web-scraped corpora with its initial number of unclassified texts equal to 37260 (Table 3.3) and our area of interest of ca. 2500 square km. Davies (2013) collected 14231 thematically unclassified texts for an area of 432 square km using place names from Ordnance Survey 1:100000 map as search terms and Kim et al. (2015) scraped 16527 texts for Melbourne and 590 texts for Santa Fe with no thematic classification.

One of the most important questions related to a corpus of first-person perception is who wrote down their experiences of landscape and who did not? This question can be approached from two perspectives: initial corpus design and final control of the actual content.

The selection of search terms is crucial for corpus design from both spatial and thematic points of view. In our work, the spatial distribution of the final corpus is strongly related to the spatial distribution of the initial search terms, how-

ever, we also retrieved texts for locations with no search terms provided, since descriptions often contain more than one location. This suggests that, by customising, we can ensure coverage of the area of interest, e.g., by including summits from Wainwright's book, *The Outlying Fells of Lakeland*, we can cover the area to the south of the Lake District, which was not captured by our search terms. However, from a thematic point of view, such a list will potentially not add additional groups of people and types of activities to our corpus. Solutions to this issue should be tested in further research, with a rather simple starting point being testing search terms that include locations of specific types of activities (e.g., 'River Esk'), potentially with words related to the activity itself (e.g., 'kayaking') in order to retrieve descriptions of these activities.

Analysis of the actual contributors is another important way of ensuring whose opinions are reflected. We suggest that writers, as captured by our corpus, could and should be surveyed or interviewed to reveal, e.g., their underlying motivation for contributing and their demographics [c.f. *boyd*, 2007]. However, concerning the balance of the corpus, more theoretical work has to be done to cover the specifics of this new source of data, since it is not yet well-established with which categories contributors should be analysed. In traditional approaches to landscape monitoring, information such as age, gender and occupation is considered important [*Kienast et al.*, 2012]. The expert group described in Section 3.6.5 additionally found ethnicity and mobility restrictions as crucial factors related to the Equality Act 2010 [*UK Government*, 2013]. One critic of the expert-dominance of the LCA approach, Andrew Butler, suggested other categories by adapting work that Edward Relph (1976) used for classification of place types to perception of landscapes (2016), with categories gradually changing from objective values of landscapes identified by authorities and planners to the sense of belonging typically experienced by long-term inhabitants of a landscape (Table 4.1).

We attempted to include these ideas by using two lists of search terms, which we assumed would let us extract texts of two different groups of users: 'behavioural insiders' and 'empathetic insiders'. However, we also collected many descriptions of fell running and cycling contributed by locals. Landscapes, in such cases, act as affordances allowing these types of actions, and one might then classify the contributors as 'incidental outsiders'. This makes locals a dynamic group of users, changing from 'incidental outsiders' to potentially 'existential insiders'. Additionally, the term 'local' is itself a vague category as noted in the expert discussion, since there are people who have lived a relatively long time in the region, but still consider themselves 'offcomers', because they were not born there[1]. Komossa et al. (2018) adapted modes of tourist experiences identified by

1 'For the moment I think I will concentrate on what it is like to be an 'offcomer' to the lake district. Like an insider guide (I have lived here for about sixteen years) but not written by a true insider (ie: a born and bred Cumbrian).', source: https://lifeinwindermere.blogspot.com/2008/08/raining-in-lake-district.html

Table 4.1: Categories of different groups of landscape users and their characteristics.

Category	Characteristics [*Relph*, 1976, p. 52-53]
Objective outsider	Objective or 'dispassionate' attitude towards landscapes allowing, e.g., planners to 'restructure them [landscapes] according to principles of logic, reason and efficiency'
Incidental outsider	Landscapes are a 'background or setting for activities and are quite incidental to those activities', e.g., conference participants, truck drivers.
Vicarious insider	Landscapes are experienced through 'secondhand' way, e.g., through poems, lyrics, films.
Behavioural insider	As opposed to incidental outsider, landscapes are deliberately visited, visual patterns play primary role, e.g., tourists
Empathetic insider	Landscapes are not just looked at, but their identity is appreciated with 'deliberate effort of perception' and understanding of 'place as rich in meaning', e.g., locals.
Existential insider	Sense of belonging to a landscape, e.g., long-term inhabitant.

Cohen (1979) to the following categories: convenience recreationist, day tripper, education recreationist, nature trekker and spiritual recreationist. These categories make it possible for users to move from one category to another depending on the activity, in contrast to the categories suggested by Relph (1976). However, they do not represent all the activities possible in a landscape, which was an important issue in our expert group discussion, and are limited to short-term recreation.

Additional difficulties in the creation of the corpus include assigning texts to space and chunking of these texts on comparable documents. In the georeferencing step, there are two points which still require improvement. Our method could not distinguish between locations visited or simply seen, as was also noted in the expert discussion. Such methods exist for Romance languages, but they have not been adapted to English yet [*Moncla et al.*, 2014]. In the 'contrasting sound' class of aural perception a quiet place was often contrasted to a loud or busy one in the same sentence. We did not implement a way to capture such situations; therefore, selection between two locations in the georeferencing step of this class was done manually. However, the expert group explicitly found it important that this class of descriptions is separated from the others, since

often two sound emitters are present in such descriptions (e.g., perceived silence disturbed by jet engine noise). The step of temporal document chunking with respect to the dates of individual experiences is challenging, since first-person narratives can be of different types (e.g., a week-long experience versus one climb described in one text); therefore, we suggest that here a classification scheme has to be first developed, taking into account temporal and spatial distribution of texts.

Two other issues relate to scraping and the archiving of such scraped data. Some guidelines for web-scraping already exist [e.g., *Greenaway*, 2017] and their implementation must be ensured in any project related to text scraping. Archiving of such texts is necessary for the purposes of monitoring, since webpages might cease to exist. In our study area, an important domain hosting multiple descriptions www.trekkingbritain.com is not available anymore, though we can see on the webarchive resource that it still existed on the 30th of March 2019[2]; however, in such archives, we have only snapshots of texts, which are not sufficient for landscape monitoring [*Hale et al.*, 2017]. Archiving is undoubtedly related to the ethical and legal issues of such processes, where some points are controversial. For example, to ensure data anonymity, author information must be stored separately from the texts, however, to give credit for the author's work as under Creative Commons license, author information must be provided [e.g., *Zimmer*, 2018].

Key messages:

- We developed a reproducible workflow allowing for the creation of a corpus of ca. 7000 rich, spatially and thematically relevant texts describing first-person perception in the Lake District.
- Spatial and thematic properties of the search terms are replicated in the collected texts.
- The criteria to assess how balanced a corpus is need further development.
- Methods to automatically distinguish between locations that have been visited or simply seen should be further refined.
- Reproducibility remains a problem for both legal and ethical reasons.

4.2 VARIATION OF PERCEIVED LANDSCAPE PROPERTIES AS EXTRACTED FROM TEXT CORPORA (RQ2, RQ3)

To explore variation in landscape properties, we created a workflow to extract relevant texts from our corpora and classify them according to properties of visual, aural and olfactory perception, as well as tranquillity. We tested our methods on different collections of texts; namely, descriptions of photographs

2 https://web.archive.org/web/20190330150744/http://www.trekkingbritain.com/

contributed to the project Geograph, historical texts collected in the CLDW and our corpus of first-person perception in the Lake District.

For visual perception, we demonstrated that language allows the modelling of scenicness, which, on the one hand, legitimise our further work with textual sources, and on the other, opens new possibilities for regions where the ScenicOrNot dataset does not exist, but where rich textual descriptions can be collected. Transferring such a model to another region, however, would require taking into account local perception of landscapes [*Hull IV and Reveli*, 1989; *Hägerhäll et al.*, 2018], differences in ontologies of landscape [*Comber et al.*, 2005] and issues related to translation of concepts [*Burenhult et al.*, 2017].

Extraction of descriptions referring to the most scenic and the most unattractive landscapes was carried out by building a domain-specific lexicon, where additional clues were used in the form of ScenicOrNot votes for assigning polarity [*Kaji and Kitsuregawa*, 2007; *Lu et al.*, 2011]. Our results demonstrated that individual words are not always rated as expected in landscape research, with the most prominent example being the word 'water', rated on average between 3 and 5 out of 10, despite being commonly associated with preferred landscapes [*Yang and Brown*, 1992; *Schirpke et al.*, 2013]. On closer examination, 'water' is revealed to be a component of rural landscapes, as opposed to perceived natural scenes (Figure 3.14). This finding made it important to take into account the context in which landscape elements are perceived. Methodologically, we followed this up in two ways; by taking a description as a whole in the case of tranquillity, aural and olfactory perception, and by going beyond a bag-of-words model in visual perception to model the dependencies between words [*Manning and Schutze*, 1999]. However, the rating of photographs was based purely on the visual components, whereas the descriptions contain important concepts, as broadly discussed in this work, such as references to other types of perception, including experiential concepts such as 'steep ascent'. This shortcoming suggests that ratings for such syntactic pairs have to be collected in a separate experiment, where the demographics of annotators will be gathered in parallel, since we expect landscape preference to vary geographically and culturally [*Hull IV and Reveli*, 1989; *Hägerhäll et al.*, 2018].

Our contributions in the exploration of sound experiences as described in written texts are threefold. First, we proved the usefulness of the sound emitter taxonomy used in ecoacoustics for human-centred approaches [*Krause*, 2008; *Pijanowski et al.*, 2011b], and developed a classification for tranquillity through macro- and micro-analysis, reflecting the complexity and richness of natural language as opposed to continuous scale from the least to the most tranquil areas [*MacFarlane et al.*, 2004; *Hewlett et al.*, 2017]. Second, we developed heuristics allowing texts to be extracted related to aural perception and tranquillity with a precision of 0.75. However, transfering to a new corpus introduced noise, which we had to identify and control heuristically. For example, in the Geograph corpus, one of the false positives was 'echo', used as verb (Section 3.5.2),

whereas in the corpus of first-person perception in the Lake District the activity of running and cycling over passes introduced the metaphorical 'screaming calves' (muscles), which had to be identified and filtered out. Lastly, we trained a model that classified descriptions in both corpora according to emitters such as biophony, geophony, anthrophony and absence of sound with a precision of 0.81. This classification revealed that perceived sound emitters are present in nearly every description of sound experiences [*Fisher*, 1999; *Morton*, 2009]. Thus, we created a classified corpus of sound descriptions consisting of about 11000 descriptions, ca. 9000 of which are associated with coordinates of photographs and the other 2000 are associated with a location extracted from the description itself. This corpus can be used as training data in future classification tasks of aural perception.

As opposed to the domain-specific lexicon that we created for visual perception, we used a general lexicon in the case of sentiment classification for aural perception. The majority of the words contained in our descriptions (ca. 93%) were not present in the lexicon and their values had to be assigned using word embeddings pre-trained on a corpus of general, as opposed to domain-specific, texts of the Common Crawl corpus (Section 2.3.1). However, even this simplistic approach to sentiment analysis demonstrated that natural sounds do not necessarily have positive connotations [*Fisher*, 1999]. Zooming in to individual classes of sound, we could see that, for example, in biophony, sounds associated with domesticated animals, as opposed to wild, are perceived less positively. In geophony, references to adverse weather have negative connotations, as opposed to sounds of sublime waterfalls and beautifully murmuring streams. In anthrophony, descriptions that can be potentially classified as technophony [*Mullet et al.*, 2016] are mostly related to negative sentiments, whereas other types of anthrophony are perceived more positively.

For olfactory perception, despite the small number of retrieved texts, we were able to observe important spatial patterns, revealing an additional experiential dimension of landscapes. However, analysis of the granularity of areas of distinctive character or individual landscape elements seems to be less suitable for this type of perception, and we suggest, instead, building spatial clusters without predefined borders [e.g., *Gao et al.*, 2017], by, for example, building a heather-smelling cluster on the 'covered slopes' of Skiddaw and Blencathra. Testing different space partitions, such as spatial clusters, grid-tessellations or approaches based on identification of landforms (e.g., valleys and mountains) could be potentially useful for other types of perception [*Fisher et al.*, 2004; *Hobel et al.*, 2016; *Jenkins et al.*, 2016], especially in the step of spatial comparison of landscapes, described in the following section.

Key messages:

- Our methods of extraction and classification of different types of perception were tested on three corpora: Geograph descriptions, CLDW and the corpus of first-person landscape perception in the Lake District created in this work.
- We have created a domain-specific lexicon for visual perception, going beyond a simple bag-of-words model.
- Taxonomies of aural and olfactory perception proved to be useful for a human-centred approach. Our taxonomy of tranquillity, developed by a combination of macro- and micro-analysis, unveiled new insights into landscape perception.
- Sentiment analysis revealed that natural sounds do not always have positive connotations and that descriptions of tranquillity are overall more positive than the corpora from which they are extracted.
- Other approaches to space partitions could be tested for better understanding of spatial properties of different types of perception.

4.3 CHARACTERISATION AND COMPARISON OF LANDSCAPES BASED ON TEXTUAL DESCRIPTIONS (RQ4)

Our work has demonstrated that written first-person narratives can be used to characterise changes in perceived landscape properties across both space and time. Changes in space were presented at four levels of granularity: Great Britain, the Lake District, areas of distinctive character and individual landscape elements. Since we have classified the descriptions and assigned them to space, we were able to perform both micro- and macro-analysis by zooming in to individual descriptions in the former and by analysing the number of descriptions of a particular class per area of distinctive character or per type of landscape element in the latter.

At the level of Great Britain, we were able to extract the overall distribution of scenicness and descriptions of aural perception, which allowed us to compare such overall patterns with, for example, patterns distinctive within the boundaries of the UK's 15 national parks. At the level of the Lake District, one of the interesting results is that the tranquillity class 'no-movement' mostly describes mirror-like reflections on the lakes and tarns and, thus, characterises the Lake District as a whole, as influenced by the Romantic poets [*Selman and Swanwick*, 2010], rather than individual locations. From the quantitative point of view, we see that this region is dominated by syntactic pairs describing scenic characteristics of a landscape and by descriptions of perceived tranquillity. Analysis on the level of areas of distinctive character showed that, for example, in the

areas of 'Scafell Massif' and 'Skiddaw and Blencathra', visual descriptions are uniquely positive, whereas sound experiences include negative references to anthrophony. The area of 'Upper Windermere', however, has negative anthropogenic influence in both visual and aural perception. We were also able to compare all areas to each other, for example, based on the normalised number of descriptions related to different classes of aural perception and tranquillity, where the aforementioned 'Upper Windermere' is the area with the highest proportion of anthrophony, and areas south of Ullswater, characterised by the sounds of roaring stags (Figure 3.22), exhibit the highest proportion of biophony. We further characterised and compared individual landscape elements, where the sublime waterfall Aira Force and beautiful Ashness Bridge are both in the ten most frequently mentioned locations and are both characterised by geophony. However, zooming in to their descriptions revealed that, as in the case of visual perception, Aira Force is characterised by the sublime sounds of 'thundering' and 'deafening' water, whereas Ashness Bridge is over a stream emitting 'wonderful' sounds (Section 3.6.4).

Comparing landscapes over time based on written first-person narratives presents three main challenges, since, first, language itself varies diachronically [*Williams*, 1976]. Second, language reflects physical changes in landscapes, and, lastly, it also contains changes of landscape perception [*Gregory et al.*, 2015]. These three aspects are strongly intertwined, and to unravel this tangle, we explored descriptions of tranquillity. For example, diachronic change was revealed by different proportions of presence of our silence-related seed words in the historical CLDW (89%) and in the modern Geograph corpora (10%). Silence is not only more common in the CLDW but is also expressed with greater linguistic variety and higher ambiguity of individual words, such as 'quiet' and 'peace'.

Historical change of landscape was particularly visible through co-occurrences with our seed words and through mapping of the corpora. Transport related words (e.g., 'road', 'motorway') and fine granularity generic locations (e.g., 'spot', 'corner') emerged in the modern corpus and their mapping revealed that they are mostly located near transport arteries, demonstrating that modern authors can find tranquil locations in a potentially unpromising settings and that total silence is not necessary for the discovery of tranquillity. This observation is supported by the disappearance of descriptions of classified total silence over time and increased importance of silence, expressed by contrasting it to other sounds, especially of anthropogenic nature related to increased visitor numbers, transport infrastructure and other signs of the landscape's commodification[3] [*Pheasant and Watts*, 2015].

From the point of view of changing perception, silence and quietness have undergone a profound change. Before the 18th century, and thus before the ma-

3 'Right by the A66, but quiet nevertheless', source: https://www.geograph.org.uk/photo/4325839

jority of the CLDW texts were written, quietness was perceived as a lack of civilisation [*Fisher*, 1999], which shifted over time to appreciation of tranquillity, when industrialisation and the noise it brought with it made people pursue silence and tranquillity [*Taylor et al.*, 2018]. Contemporary authors maintain this attitude towards tranquillity and search for it in rural landscapes and countryside [*MacFarlane et al.*, 2004]. However, what is meant by quiet and tranquil has changed in the time between writing of the texts in the two corpora. Micro-analysis of the texts demonstrated that authors of the historical corpus meant two things when using these words; peacefulness in the landscape, as well as mental calm, and felt obliged to explain this link between physical properties of the landscape and their mental state, whereas modern writers take this connection for granted. This observation, on the one hand, brings us back to diachronic variation of language, since historical authors need more diverse language to describe their sense of tranquillity. On the other hand, they influenced the overall positiveness and necessity of tranquillity as perceived today, which we can also see by mapping our corpora (Figure 3.21), where the influence of Wordsworth's writing becomes clear through the emergence of a tranquillity cluster next to Grasmere, a region where Wordsworth resided for more than fifty years and which became popular in search of the sounds of Wordsworthshire [*Donaldson et al.*, 2015].

Key messages:

- Our approach allows us to perform both micro-analysis, by zooming in to individual descriptions, and macro-analysis by comparing numbers of descriptions of a particular class, e.g., per area of distinctive character or per type of landscape element.
- In the comparison of landscapes over time, using our approach, it is possible to detect diachronic variation of language, physical changes of landscapes and changes in landcsape perception.

4.4 TOWARDS LANDSCAPE MONITORING AND LANDSCAPE ASSESSMENT (RQ5)

Our work demonstrated that written first-person narratives allow the exploration of patterns of perceived landscape properties globally and locally. We can compare the Lake District as a whole to, for example, Areas of Outstanding National Beauty, highlighting particularities of each and identifying differences. On the granularity of areas of distinctive character, for example, more and less tranquil areas can be identified and, on the level of individual landscape elements, potentially disturbing or attracting factors. Thus, we can extend insights

revealed through spatial patterns of visible features of landscapes to include tranquillity, and aural and olfactory characteristics.

The expert group identified the value of our approach to LCA and for other activities of the National Park Authorities, such as, for example, exploration of sentiments towards measures related to access management, confirming that the approach has practical utility. We also see great potential for written first-person narratives in landscape monitoring in cases where the data preceding an event causing a potential change has to be collected. For example, one of the tranquillity descriptions we extracted is about a newly listed mountain in the Lake District called Miller Moss. After being resurveyed this mountain was measured higher than 2000 feet. In our corpus authors write that 'there is tranquillity here compared with much busier fells of Blencathra and Skiddaw' and that 'Miller Moss is not the most exciting hill in the world but it should become a little busier now'[4] since it was added to a Nuttall hill-bagging list based on these new measurements. Our methods can be useful for such monitoring, because, by scraping texts from the internet we can 'go back in time' and check if the tranquillity of a location has changed since it was added to a hill-bagging list. This relates to the granularity level of individual landscape elements. On the level of areas of distinctive character, we also see this potential. As described in Section 3.1, the National Trust has bought a piece of land from where J.M.W. Turner painted a panorama in 1797. It is very likely that this quiet part of the Lakes (Lorton Vale) will receive more visitors due to this recent media interest and its new owner. Another example could be complementary fine-grained information for GIS modelling on a global scale. In our work on tranquillity exploration, we were able to demonstrate that GIS methods based on proximity to potential noise emitters as a proxy of disturbance [*Carver et al.*, 2002; *MacFarlane et al.*, 2004] do not model tranquillity pockets close to transportation arteries, which are potentially even more important to people's well-being, since they are more accessible.

In Section 2.1.3 we briefly mentioned the method of Historical Landscape Characterisation (HLC), which was developed to provide a historical understanding and perspective to LCA [*Fairclough and Herring*, 2016]. HLC categories claim to be objective, such as 'late 19th century woodland', 'abandoned industry', extracted using archaeological approaches, remote sensing imagery and historical map materials. Here we also see a potential for our approach as an additional source of not only perceptual, but also factual information, where basic spatial data in the form of cartographic products or aerial photographs does not exist. However, especially the perceptual component can be found in written first-person narratives; even the HLC pioneers themselves emphasise the importance of memories and historical relationships with place by giving the example of Thomas Hardy, for whom a certain crossroads near Boscastle, where he had

4 https://www.grough.co.uk/magazine/2018/08/09/england-has-a-new-mountain-miller-moss-now-go-find-it

fallen in love, is particularly important and reminds him of his lost love. Hardy immortalized this information in a poem, and this source of information gives an additional glimpse of the history of this place, since 'history, archaeology, poetry, memory and associations all turn Boscastle into cultural landscape, and all contribute to the ways people value it.' [*Fairclough and Herring*, 2016, p. 192]. Similar observations are made by Jess Edwards (2018), who advocates using existing literature and supporting creative writing of landscapes' users for LCA.

Key messages:

- Written first-person narratives as a data source and our methodological workflows were positively rated by a local expert group in the Lake District.
- A single dataset, and thus a single method, can hardly characterise all aspects of a landscape, and we see the solution in a combination of different approaches, where our approach shows potential in demonstrating global patterns, and at the same time as a complementary source of fine-grained information to, for example, GIS modelling.
- Our approach can be used in different landscape assessment programs, including the broadly discussed in this work LCA, but also in HLC, as an additional source to fill data gaps and add a perceptual layer of information.

4.5 OVERALL LIMITATIONS

Our approach has several limitations, including those related to data sources and methods. Two main obstacles are that, first, not everybody chooses to write down their experiences [*Michelin et al.*, 2011] and, second, human language has a positivity bias [*Dodds et al.*, 2014]. These issues are less relevant for the Geograph corpus, where the descriptions are shorter and defined by the conditions of the game, which additionally encourages descriptions of less popular landscapes through a point-system. As we described in detail in Section 4.1, we do not know the underlying demographics or motivation of the writers, making a study revealing this information an important step for the future endeavours. In the case of Geograph, however, some information is available through an anonymous survey conducted by the initiators of the project.

Another potential source of uncertainties is the understanding of the scale in questions using Likert scale, which is relevant for the ScenicOrNot dataset, e.g., 'is 10 or 1 the most scenic?' Such questionnaires can be also prone to vandalism. In one of our examples the word 'traffic' had the following ratings '9,1,1,2,6,1,2,2,2,1,2,2,5', where the rating 9 could be potentially a result of misunderstanding or vandalism [*Neis et al.*, 2012].

A second group of limitations relates to the methodology. Despite advances in NLP, it is still challenging to accurately analyse natural language. State of the art tools are robust but are not flawless as we saw in the example in Section 2.3.3, where 'interrupted' was identified as an adjective instead of a past participle verb. Another practically unsolved problem in NLP is the plethora of metaphors in language [*Liu*, 2012], especially relevant in our case for descriptions of tranquillity, where some of the cases were highly problematic, even for human annotators.

To extract descriptions, we, first, privileged precision over recall and, second, only used keywords that were expected to lead us to the desired descriptions. However, it is possible that there are more descriptions which we did not foresee in the data. To overcome this issue, we can extend these search words by using pre-trained word embeddings or use all our annotated and extracted descriptions as training data to search for more relevant descriptions. This data would reflect a 'positive' class of the training data, meaning that such descriptions are what we are looking for. In the future work, more research has to be done to create an appropriate 'negative' class to indicate which descriptions we are not interested in. Training word embeddings ourselves, as well as using other methods based on artificial neural networks, would require large annotated datasets. For example, the GloVe word embeddings used in this work are trained on 42 billion tokens from the Common Crawl corpus [*Pennington et al.*, 2014].

Lastly, it is crucial to estimate the possible harm of such seemingly innocuous applications. For example, if we publish a map of tranquil places publicly, this information may potentially lead to the overcrowding of such locations [*Zimmer*, 2018].

4.6 NEW POSSIBILITIES AND WAYS FORWARD

Important sources of bottom-up spatial information, such as participatory mapping and crowdsourcing, are rapidly developing tools to increase participation through online hosting and gamification [*Bayas et al.*, 2016; *Gottwald et al.*, 2016; *Baer et al.*, 2019; *Bubalo et al.*, 2019]. We suggest that our approach adds to this tendency with the possibility of collecting data for large spatial extents, thus, joining Hu (2018) in his calls for broader use of texts in GIScience and for more thematically focused research rather than attempts to customise existing methods for methodologically interesting questions with limited domain relevance. In this work we showed the potential of applying texts and a combination of GIScience and NLP methods to the domain of landscape characterisation. In this section we suggest other questions that can be tackled with a similar approach.

We have not explored memories and associations as another factor of perceived landscape properties as identified in the LCA guidelines (Figure 2.1), but through micro-analysis of corpora used in this work we know that there are references to both cultural (e.g., myths and wars) and personal memories. We see great potential in our approach to explore these aspects of landscape perception together with wilderness, another landscape-relevant concept, which has been demonstrated to be related rather to a feeling of remoteness, than to its physical measure [*Pheasant and Watts*, 2015] and, thus, particularly hard to model.

We discussed the importance of the Picturesque movement on the landscape conservation in Britain and on the ways landscapes were – and are – perceived. The terminology used to describe aesthetic landscapes (e.g., 'sublime', 'picturesque') was analysed in the historical corpus CLDW by Donaldson et al. (2017). We suggest that our corpus of first-person landscape perception in the Lake District and a more detailed approach to georeferencing can reveal new insights in the ways 'language of landscape appreciation' has developed (or regressed?) with time [*Donaldson et al.*, 2017]. Our workflows to extract and classify sound and smell experiences can be used directly in the field of digital humanities or tourism studies to extract references to sounds and smells from fiction [e.g., *Dann and Jacobsen*, 2003; *Buciek*, 2019]. Comparing our findings with the results of psycholinguistics can give additional insights into the ways perception is conceptualised and expressed through language [e.g., *Majid and Levinson*, 2011].

Not only the 18th-19th century theories on visual perception and aesthetics have influenced the ways landscapes are perceived today. The writings of local authors, such as Alfred Wainwright, also have an impact on which landscapes are visited and what is written about them [*Palmer and Brady*, 2007]. In the *Pictorial Guides to the Lakeland Fells*, each of 214 fells is described together with sketches, panoramas, routes of ascent and descent, and most importantly, with a plethora of subjective observations and opinions from Wainwright himself (Figure 3.2). In the personal notes in the conclusion of volume 7, Wainwright published the list of the finest summits of the Lake District: Scafell Pike, Bowfell, Pillar, Great Gable, Blencathra and Crinkle Crags [*Wainwright*, 1966]. His favourite fell Haystack, despite being 'the best fell-top of all – a place of great charm and fairyland attractiveness', did not make it in this list, because of its 'inferior height' [*Wainwright*, 1966, p. Haystacks 10], since his criteria were 'height, a commanding appearance, a good view, steepness and ruggedness' [*Wainwright*, 1966]. These criteria are clearly related to the concepts of sublime and picturesque. However, he added another dimension to landscape perception: the one of the fell-walker and not merely of the fell-observer, namely the dimension of the 'bodily experience' [*Palmer and Brady*, 2007, p. 405]. For him, marches and moors were less aesthetic not only because of their uniformity, but also because of the way they impede walking. We did not look at this experiential dimension in detail in our

work, but we see some clear potential in our workflows and our newly created corpus to explore it further, since Wainwright did not only describe his favourite fells, but also the ones he found boring or ugly. For example, Mungrisdale Common 'from whatever side it is seen, has no more pretension to elegance than a pudding that has been sat on.' [*Wainwright*, 1962, p. 1 Mungrisdale Common]. Wainwright himself did not anticipate that just by writing about this fell, even in the negative tones, he put it on a hill-bagging list and made large numbers of people visit it. As one author in our corpus writes, 'So I had to go, purely because Wainwright had put it in his book. In a book where he'd told me not to go there.'[5] Therefore, we get a higher variety of landscape descriptions, than just the 'best' ones, making it possible to explore different aspects of Wainwright's influence on the Lake District perception.

Finally, we used a variety of methods to present our results: grid-tessellations for overall spatial distribution patterns, word clouds for semantic information within these grids, larger regions or areas of distinctive character, and point locations for individual descriptions. Our choice of word clouds was motivated by the need to demonstrate a large number of words/ word-pairs in a user-friendly way as opposed to, for example, sophisticated but rather hard to interpret spatial treemaps [*Purves et al.*, 2011]. However, more than 150 words or 100 word-pairs are hard to visualise using word clouds, and their other limitations include unclear focus (which words are noticed first?) and comparability issues (does everybody notice the same words?). We georeferenced aural perception, tranquillity and olfactory perception as points and visualised them using symbols suggesting vagueness (e.g., Figure 3.22). However, apart from georeferencing issues covered broadly in previous sections, we suggest that, for visualisation purposes, the objects that emit sounds or smells should be further classified, for example, into dynamic/ static (e.g., cars/ waterfalls) or point-wise/ distributed (e.g., restaurant/ blossoming flowers) and be visualised correspondingly in order to enhance communication of landscape properties for the LCA users.

5 https://ramblingman.org.uk/walks/wainwrights/northern-fells/mungrisdale-common

5

CONCLUSIONS AND OUTLOOK

In this work we set out to explore the potential of written first-person narratives for landscape monitoring and assessment, since we hypothesised that such texts reflect natural ways of humans to interact about locations and are rich on perceptual information. By using a georeferenced corpus and landscape preference ratings for the extent of Great Britain, we were able to explain 52% of the variability in visual landscape preference modelled as a function of language, suggesting the potential of further work with textual data sources. We then moved from a pragmatic way of selecting data sources, motivated by the ease of access or processing (e.g., only georeferenced data), to creating a thematically and spatially relevant corpus of first-person landscape perception that contains more than 8 million of words. This step is crucial for analysing non-urban landscapes, where well-distributed coverage is not guaranteed through projects such as Geograph and for analysing locations, which are not mapped through other data sources, (e.g., not *Instagrammable* landscapes).

Initial experiments showed that sound-related experiences could be extracted from written first-person narratives and classified based on sound emitter (e.g., crashing waves, singing birds) as suggested in the field of ecoacoustics. In an interdisciplinary study we explored diachronic variations in references to sounds and their absence in the English Lake District using not only macro- but also micro-analysis to understand the data and its underlying patterns, an often forgotten analysis task in the 'big data' era. This combination allowed us to develop a tranquillity taxonomy, reflecting the particularities of this concept as expressed through natural language. Thus, using existing building blocks from GIScience and NLP, we created repeatable workflows from the selection of one of the senses, through extraction and classification, to a map of its spatial distribution. In a final study we integrated these building blocks to demonstrate the practical utility of written first-person narratives and our processing workflows by applying them directly to LCA.

In this work we advocated for a combination of different approaches and different data sources, since no single method or dataset can cover all groups of users and be available for every territory. Therefore, further research should focus on finding suitable ways to integrate heterogeneous data, such as expert evaluations of landscapes, opinions present in written first-person narratives, coded interviews, lists available through free-listing, spatial outputs of GIS modelling or PPGIS, and continuous scales of survey's outputs.

REFERENCES

Acheson, E., S. De Sabbata, and R. S. Purves, A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, *Computers, Environment and Urban Systems*, *64*, 309–320, doi: 10.1016/j.compenvurbsys.2017.03.007, 2017.

Adams, B., and M. Gahegan, Exploratory Chronotopic Data Analysis, in *Geographic Information Science. GIScience 2016*, Montreal, Canada, doi: 10.1007/978-3-319-45738-3, 2016.

Adams, B., and G. McKenzie, Inferring Thematic Places from Spatially Referenced Natural Language Descriptions, *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, pp. 201–221, doi: 10.1007/978-94-007-4587-2, 2013.

Agarwal, S., and H. Yu, Automatically Classifying Sentences in Full-Text Biomedical Articles into Introduction, Methods, Results and Discussion, *Bioinformatics*, *25*, 3174–3180, 2009.

Agnew, V., Hearing Things: Music and Sounds the Traveller Heard and Didn't Hear on the Grand Tour, *Cultural Studies Review*, *18*(3), 67–84, doi: 10.5130/csr.v18i3.2855, 2012.

Ahern, S., M. Naaman, R. Nair, and J. H. I. Yang, World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections, in *Proceedings of the ACM International Conference on Digital Libraries*, pp. 1–10, Vancouver, British Columbia, Canada, doi: 10.1145/1255175.1255177, 2007.

Aiello, L. M., R. Schifanella, D. Quercia, and F. Aletta, Chatty maps: constructing sound maps of urban areas from social media data, *Royal Society Open Science*, *3*, 150,690, doi: 10.1098/rsos.150690, 2016.

Amitay, E., N. Har'El, R. Sivan, and A. Soffer, Web-a-where: geotagging web content, in *Proceedings of SIGIR '04 conference on Research and development in information retrieval*, pp. 273–280, Sheffield, South Yorkshire, UK, doi: 10.1145/1008992.1009040, 2004.

Baer, M. F., F. M. Wartmann, and R. S. Purves, StarBorn: Towards making in-situ land cover data generation fun with a location-based game, *Transactions in GIS*, pp. 1–21, doi: 10.1111/tgis.12543, 2019.

Ballatore, A., and B. Adams, Extracting Place Emotions from Travel Blogs, in *AGILE 2015, Lecture Notes in Geoinformation and Cartography*, pp. 1–5, Lisbon, Portugal, 2015.

Baroni, M., and S. Bernardini, BootCaT: Bootstrapping corpora and terms from the web, in *Proceedings of LREC*, vol. 4, pp. 1313–1316, Lisbon, Portugal, 2004.

Bayas, J. C. L., et al., Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology, *Remote Sensing*, *8*(11), 905, doi: 10.3390/rs8110905, 2016.

Benfield, J. A., P. A. Bell, L. J. Troup, and N. C. Soderstrom, Aesthetic and affective effects of vocal and traffic noise on natural landscape assessment, *Journal of Environmental Psychology*, *30*(1), 103–111, doi: 10.1016/j.jenvp.2009.10.002, 2010.

Benson, J., Aesthetic and Other Values in the Rural Landscape, *Environmental Values*, *17*, 221–238, doi: 10.3197/096327108X303864, 2008.

Bieling, C., Cultural ecosystem services as revealed through short stories from residents of the Swabian Alb (Germany), *Ecosystem Services*, *8*, 207–215, doi: 10.1016/j.ecoser.2014.04.002, 2014.

Bieling, C., T. Plieninger, H. Pirker, and C. R. Vogl, Linkages between landscapes and human well-being: An empirical exploration with short interviews, *Ecological Economics*, *105*, 19–30, doi: 10.1016/j.ecolecon.2014.05.013, 2014.

boyd, d., Why youth (heart) Social network sites: the role of networked publics in teenage social life, in *MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media Volume*, edited by D. Buckingham, MIT Press, Cambridge, MA, doi: 10.1162/dmal.9780262524834.119, 2007.

boyd, d., and K. Crawford, Six Provocations for Big Data, *Computer*, *123*(1), 1–17, doi: 10.2139/ssrn.1926431, 2011.

Brady, E., *Aesthetics of the natural environment*, 287 pp., University of Alabama Press, Tuscaloosa, 2003.

Brown, G., and P. Reed, Public Participation GIS: A New Method for Use in National Forest Planning, *Forest Science*, *55*(2), 166–182, 2009.

Bruns, D., and B. Stemmer, Landscape Assessment in Germany, in *Routledge Handbook of Landscape Character Assessment*, pp. 154–167, Routledge, 2018.

Bubalo, M., B. T. V. Zanten, and P. H. Verburg, Crowdsourcing geo-information on landscape perceptions and preferences: A review, *Landscape and Urban Planning*, *184*, 101–111, doi: 10.1016/j.landurbplan.2019.01.001, 2019.

Buciek, K., Soundscape and Heritage: The Sonic Environment in Roskilde Juxtaposed with James Joyce's Ulysses, *GeoHumanities*, *5*(1), 86–102, doi: 10.1080/2373566x.2018.1531720, 2019.

Burenhult, N., and S. C. Levinson, Language and landscape: a cross-linguistic perspective, *Language Sciences*, *30*(2-3), 135–150, doi: 10.1016/j.langsci.2006.12.028, 2008.

Burenhult, N., C. Hill, J. Huber, S. van Putten, K. Rybka, and L. San Roque, Forests: the cross-linguistic perspective, *Geographica Helvetica*, *72*(4), 455–455, doi: 10.5194/gh-72-455-2017, 2017.

Buscaldi, D., and B. Magnini, Grounding Toponyms in an Italian Local News Corpus, in *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10*, Zurich, Switzerland, doi: 10.1145/1722080.1722099, 2010.

Butler, A., Dynamics of integrating landscape values in landscape character assessment: the hidden dominance of the objective outsider, *Landscape Research*, *41*(2), 239–252, doi: 10.1080/01426397.2015.1135315, 2016.

Butler, J. O., C. E. Donaldson, J. E. Taylor, and I. N. Gregory, Alts, Abbreviations, and AKAs: Historical onomastic variation and automated named entity recognition, *Journal of Maps and Geography Libraries*, *13*(1), 58–81, doi: 10.1080/15420353.2017.1307304, 2017.

Carles, J. L., I. L. Barrio, and J. V. De Lucio, Sound influence on landscape values, *Landscape and Urban Planning*, *43*(4), 191–200, doi: 10.1016/S0169-2046(98)00112-1, 1999.

Carver, S., A. Evans, and S. Fritz, Wilderness Attribute Mapping in the United Kingdom, *International Journal of Wilderness*, *8*(1), 24–29, 2002.

Casalegno, S., R. Inger, C. DeSilvey, and K. J. Gaston, Spatial Covariance between Aesthetic Value & Other Ecosystem Services, *PLOS ONE*, *8*(6), doi: 10.1371/journal.pone.0068437, 2013.

Caspersen, O. H., Public participation in strengthening cultural heritage: The role of landscape character assessment in Denmark, *Geografisk Tidsskrift – Danish Journal of Geography*, *109*(1), 33–45, doi: 10.1080/00167223.2009.10649594, 2009.

Chen, Y., J. R. Parkins, and K. Sherren, Using geo-tagged Instagram posts to reveal landscape values around current and proposed hydroelectric dams and their reservoirs, *Landscape and Urban Planning*, *170*(February), 283–292, doi: 10.1016/j.landurbplan.2017.07.004, 2018.

Chesnokova, O., and R. S. Purves, From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text, *Applied Geography*, *93*, 103–111, doi: 10.1016/j.apgeog.2018.02.014, 2018.

Chesnokova, O., M. Nowak, and R. S. Purves, A crowdsourced model of landscape preference, in *13th International Conference on Spatial Information Theory (COSIT 2017)*, edited by E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore, 19, pp. 19:1–19:13, Leibniz International Proceedings in Informatics, L'Aquila, Italy, 2017.

Chesnokova, O., J. E. Taylor, I. N. Gregory, and R. S. Purves, Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District, *International Journal of Geographical Information Science*, *33*, 2430–2454, doi: 10.1080/13658816.2018.1552789, 2019.

Choi, Y., and C. Cardie, Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification, in *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2, pp. 590–598, Singapore, doi: 10.3115/1699571.1699590, 2009.

Clemetsen, M., E. Krogh, and K. H. Thorén, Landscape Perception Through Participation: Developing New Tools for Landscape Analysis in Local Planning Processes in Norway, in *The European Landscape Convention. Challenges of Participation*, edited by M. Jones and M. Stenseke, pp. 219–237, Springer, doi: 10.1007/978-90-481-9932-7_11, 2011.

Coates, P. A., The Strange Stillness of the Past: Toward an Environmental History of Sound and Noise, *Environmental History*, *10*(4), 636–665, 2005.

Coeterier, J., Dominant attributes in the perception and evaluation of the Dutch landscape, *Landscape and Urban Planning*, *34*(1), 27–44, doi: 10.1016/0169-2046(95)00204-9, 1996.

Cohen, E., A Phenomenology of tourist experiences, *Sociology*, *13*, 1979.

Comber, A., P. Fisher, and R. Wadsworth, What is Land Cover?, *Environment and Planning B: Planning and Design*, *32*(2), 199–209, doi: 10.1068/b31135, 2005.

Corbin, A., *The Foul and the Fragrant: Odor and the French Social Imagination*, Harvard University Press, USA, 1986.

Council of Europe, European Landscape Convention, *Report and Convention Florence, ETS No. 17*(176), 8, 2000.

Criminisi, A., J. Shotton, and E. Konukoglu, Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, *Foundations and Trends® in Computer Graphics and Vision*, *7*(2-3), 81–227, doi: 10.1561/0600000035, 2011.

Crocker, C., and G. Jackson, Hill Bagging: the online version of the Database of British and Irish Hills, http://www.hill-bagging.co.uk/glossary.php, [Online; accessed 27-August-2019], 2001.

Dann, G. M., and J. K. S. Jacobsen, Tourism smellscapes, *Tourism Geographies*, *5*(1), 3–25, doi: 10.1080/1461668032000034033, 2003.

Daume, S., M. Albert, and K. von Gadow, Forest monitoring and social media - Complementary data sources for ecosystem surveillance?, *Forest Ecology and Management*, *316*, 9–20, doi: 10.1016/j.foreco.2013.09.004, 2014.

Davies, C., Reading geography between the lines: Extracting local place knowledge from text, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8116 LNCS*, 320–337, doi: 10.1007/978-3-319-01790-7-18, 2013.

Derungs, C., and R. S. Purves, From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus, *International Journal of Geographical Information Science*, *28*(6), 1272–1293, doi: 10.1080/13658816.2013.772184, 2014.

Derungs, C., and R. S. Purves, Mining Nearness Relations from a N-Grams Web Corpus in Geographical Space, *Spatial Cognition & Computation*, *16*(4), 301–322, doi: 10.1080/13875868.2016.1246553, 2016a.

Derungs, C., and R. S. Purves, Characterising landscape variation through spatial folksonomies, *Applied Geography*, *75*, 60 – 70, doi: 10.1016/j.apgeog.2016.08.005, 2016b.

Dodds, P. S., et al., Human language reveals a universal positivity bias, *Proceedings of the National Academy of Sciences*, *112*(8), 2389–2394, doi: 10.1073/pnas.1411678112, 2014.

Donaldson, C., I. Gregory, and P. Murrieta-Flores, Mapping 'Wordsworthshire': A GIS Study of Literary Tourism in Victorian Lakeland, *Journal of Victorian Culture*, *20*(3), 287–307, doi: 10.1080/13555502.2015.1058089, 2015.

Donaldson, C., I. N. Gregory, and J. E. Taylor, Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the English Lake District, *Journal of Historical Geography*, *56*, 43–60, doi: 10.1016/j.jhg.2017.01.006, 2017.

Drymonas, E., A. Efentakis, and D. Pfoser, Opinion mapping travelblogs, in *CEUR Workshop Proceedings*, vol. 798, pp. 23–36, Bonn, Germany, 2011.

Dunkel, A., Visualizing the perceived environment using crowdsourced photo geodata, *Landscape and Urban Planning*, *142*, 173–186, doi: 10.1016/j.landurbplan.2015.02.022, 2015.

Edwardes, A. J., and R. S. Purves, A theoretical grounding for semantic descriptions of place, *Proceedings of the 7th international conference on Web and wireless geographical information systems*, pp. 106–120, doi: 10.1007/978-3-540-76925-5_8, 2007.

Edwards, J., Literature and sense of place in UK landscape strategy, *Landscape Research*, *44*(6), 659–670, doi: 10.1080/01426397.2018.1518519, 2018.

Egorova, E., T. Tenbrink, and R. S. Purves, Fictive motion in the context of mountaineering, *Spatial Cognition and Computation*, *18*(4), 259–284, doi: 10.1080/13875868.2018.1431646, 2018.

Fairclough, G., and P. Herring, Lens, mirror, window: interactions between Historic Landscape Characterisation and Landscape Character Assessment, *Landscape Research*, *41*(2), 186–198, doi: 10.1080/01426397.2015.1135318, 2016.

Fairclough, G., I. Sarlöv Herlin, and C. Swanwick (Eds.), *Routledge Handbook of Landscape Character Assessment*, Routledge, 2018.

Fellbaum, C., *WordNet: An Electronic Lexical Database*, MIT Press, MA, 1998.

Fisher, J. A., The Value of Natural Sounds, *Journal of Aesthetic Education*, *33*(3), 26–42, 1999.

Fisher, P., J. Wood, and T. Cheng, Where is Helvellyn? Fuzziness of multi-scale landscape morphometry, *Transactions of the Institute of British Geographers*, *29*(1), 106–128, doi: 10.1111/j.0020-2754.2004.00117.x, 2004.

Friends of the Lake District, History - timeline, https://www.friendsofthelakedistrict.org.uk/pages/faqs/category/history-timeline, [Online; accessed 14-August-2019], 2019.

Gablasova, D., V. Brezina, and T. McEnery, Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence, *Language Learning*, *67*(June), 155–179, doi: 10.1111/lang.12225, 2017.

Gao, S., et al., A data-synthesis-driven method for detecting and extracting vague cognitive regions, *International Journal of Geographical Information Science*, *31*(6), 1245–1271, doi: 10.1080/13658816.2016.1273357, 2017.

GeoNames, The GeoNames geographical database, https://www.geonames.org/, [Online; accessed 22-August-2019], 2019.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, Detecting influenza epidemics using search engine query data, *Nature*, *457*(7232), 1012–1014, doi: 10.1038/nature07634, 2009.

Gliozzo, G., N. Pettorelli, and M. M. Haklay, Using crowdsourced imagery to detect cultural ecosystem services: a case study in South Wales, UK, *Ecology and Society*, *21*(3), art6, doi: 10.5751/ES-08436-210306, 2016.

Gottwald, S., T. E. Laatikainen, and M. Kyttä, Exploring the usability of PPGIS among older adults: challenges and opportunities, *International Journal of Geographical Information Science*, *30*(12), 2321–2338, doi: 10.1080/13658816.2016.1170837, 2016.

Graham, M., B. Hogan, R. K. Straumann, and A. Medhat, Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty, *Annals of the Association of American Geographers*, *104*(4), 746–764, doi: 10.1080/00045608.2014.910087, 2014.

Granö, J. G., *Pure Geography*, 191 pp., John Hopkins University Press, 1997.

Greenaway, M., Web-scraping policy, https://www.ons.gov.uk/aboutus/transparencyandgovernance/lookingafterandusingdataforpublicbenefit/policies/policieswebscrapingpolicy, [Online; accessed 28-August-2019], 2017.

Gregory, I., C. Donaldson, P. Murrieta-Flores, and P. Rayson, Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research, *International Journal of Humanities and Arts Computing*, *9*(1), 1–14, doi: 10.3366/ijhac.2015.0135, 2015.

Hägerhäll, C. M., A. O. Sang, J. E. Englund, F. Ahlner, K. Rybka, J. Huber, and N. Burenhult, Do humans really prefer semi-open natural landscapes? A cross-cultural reappraisal, *Frontiers in Psychology*, *9*(MAY), 1–14, doi: 10.3389/fpsyg.2018.00822, 2018.

Haklay, M. M., Why is participation inequality important?, in *European Handbook of Crowdsourced Geographic Information*, edited by C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, pp. 35–44, Ubiquity Press, London, doi: 10.5334/bax.c, 2016.

Hale, S. A., G. Blank, and V. D. Alexander, Live versus archive: Comparing a web archive and to a population of webpages, in *The Web as History*, edited by N. Brügger and R. Schroeder, pp. 45–61, UCL Press, London, doi: 10.1080/24701475.2018.1509579, 2017.

Hall, M. M., P. D. Smart, and C. B. Jones, Interpreting spatial language in image captions, *Cognitive Processing*, *12*(1), 67–94, doi: 10.1007/s10339-010-0385-5, 2011.

Hausmann, A., T. Toivonen, R. Slotow, H. Tenkanen, A. Moilanen, V. Heikinheimo, and E. Di Minin, Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas, *Conservation Letters*, *11*(1), 1–10, doi: 10.1111/conl.12343, 2018.

Heikinheimo, V., E. D. Minin, H. Tenkanen, A. Hausmann, J. Erkkonen, and T. Toivonen, User-Generated Geographic Information for Visitor Monitoring in a National Park: A Comparison of Social Media Data and Visitor Survey, *ISPRS International Journal of Geo-Information*, *6*(3), 85, doi: 10.3390/ijgi6030085, 2017.

Herlin, I. S., Exploring the national contexts and cultural ideas that preceded the Landscape Character Assessment method in England, *Landscape Research*, *41*(2), 175–185, doi: 10.1080/01426397.2015.1135317, 2016.

Hewlett, D., L. Harding, T. Munro, A. Terradillos, and K. Wilkinson, Broadly engaging with tranquillity in protected landscapes: A matter of perspective identified in GIS, *Landscape and Urban Planning*, *158*, 185–201, doi: 10.1016/j.landurbplan.2016.11.002, 2017.

Hill, L. L., Geographic Information System, in *Encyclopedia of Database Systems*, chap. Gazetteers, pp. 1217–1218, Springer Science + Business Media, doi: 10.1007/978-0-387-39940-9_178, 2009.

Hobel, H., P. Fogliaroni, and A. U. Frank, Deriving the Geographic Footprint of Cognitive Regions, in *Lecture Notes in Geoinformation and Cartography: Geospatial Data in a Changing World*, edited by T. Sarjakoski, M. Santos, and L. Sarjakoski, pp. 67–84, Springer International Publishing Switzerland, doi: 10.1007/978-3-319-33783-8, 2016.

Hodgson, R. W., and R. L. Thayer, Implied human influence reduces landscape beauty, *Landscape Planning*, *7*(2), 171–179, doi: 10.1016/0304-3924(80)90014-3, 1980.

Hollenstein, L., and R. Purves, Exploring place through user-generated content: Using Flickr to describe city cores, *Journal of Spatial Information Science*, *1*(1), 21–48, doi: 10.5311/JOSIS.2010.1.3, 2010.

Hu, M., and B. Liu, Mining and summarizing customer reviews, in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, vol. 04, p. 168, Seattle, Washington, USA, doi: 10.1145/1014052.1014073, 2004.

Hu, Y., Geo-text data and data-driven geospatial semantics, *Geography Compass*, *12*(11), doi: 10.1111/gec3.12404, 2018.

Hull, R., and W. Stewart, Validity of Photo-Based Scenic Beauty Judgments, *Journal of Environmental Psychology*, *12*, 101–114, doi: 10.1016/S0272-4944(05)80063-5, 1992.

Hull IV, R. B., and G. R. B. Reveli, Cross-cultural comparison of landscape scenic beauty evaluations: A case study in Bali, *Journal of Environmental Psychology*, *9*(3), 177–191, doi: 10.1016/S0272-4944(89)80033-7, 1989.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III, Deep Unordered Composition Rivals Syntactic Methods for Text Classification, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1, Beijing, China, doi: 10.3115/v1/P15-1162, 2015.

Jamie, K., *Sightlines*, Sort of Books, 2012.

Jeawak, S. S., C. B. Jones, and S. Schockaert, Using Flickr for characterizing the environment : an exploratory analysis, in *13th International Conference on Spatial Information Theory (COSIT 2017)*, 21, pp. 21:1–21:13, L'Aquila, Italy, doi: 10.4230/LIPIcs.COSIT.2017.21, 2017.

Jenkins, A., A. Croitoru, A. T. Crooks, and A. Stefanidis, Crowdsourcing a Collective Sense of Place, *Plos One*, *11*(4), 1–20, doi: 10.1371/journal.pone.0152932, 2016.

Jockers, M., *Macroanalysis: Digital methods and literary history*, University of Illinois Press, 2013.

Jones, C. B., R. S. Purves, P. D. Clough, and H. Joho, Modelling vague places with knowledge from the Web, *International Journal of Geographical Information Science*, *22*(10), 1045–1065, doi: 10.1080/13658810701850547, 2008.

Jones, M., and M. Stenseke (Eds.), *The European Landscape Convention. Challenges of Participation*, Springer, 2011.

Kaji, N., and M. Kitsuregawa, Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents., in *EMNLP-CoNLL*, vol. 43, pp. 1075–1083, Prague, 2007.

Kienast, F., B. Degenhardt, B. Weilenmann, Y. Wäger, and M. Buchecker, GIS-assisted mapping of landscape suitability for nearby recreation, *Landscape and Urban Planning*, *105*(4), 385–399, doi: 10.1016/j.landurbplan.2012.01.015, 2012.

Kienast, F., J. Frick, M. J. van Strien, and M. Hunziker, The Swiss Landscape Monitoring Program - A comprehensive indicator set to measure landscape change, *Ecological Modelling*, *295*, 136–150, doi: 10.1016/j.ecolmodel.2014.08.008, 2015.

Kienast, F., F. Wartmann, A. Zaugg, and M. Hunziker, A Review of Integrated Approaches for Landscape Monitoring, *Tech. rep.*, Report prepared in the framework of the Work Program of the Council of Europe for the implementation of the European Landscape Convention, 2019.

Kim, J., M. Vasardani, and S. Winter, Harvesting large corpora for generating place graphs, in *International Workshop on Cognitive Engineering for Spatial Information Processes*, vol. 12, Santa Fe, NM, USA, 2015.

Kisilevich, S., C. Rohrdantz, and D. Keim, "Beautiful picture of an ugly place". Exploring photo collections using opinion and sentiment analysis of user comments, *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 419–428, doi: 10.1109/IMCSIT.2010.5679726, 2010.

Koblet, O., and R. S. Purves, From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally, *Landscape and Urban Planning*, *197*, 2020.

Kolen, J., H. Renes, and K. Bosma, The Landscape Biography Approach to Landscape Characterisation, in *Routledge Handbook of Landscape Character Assessment*, pp. 168–184, Routledge, 2018.

Komossa, F., E. H. V. D. Zanden, C. J. E. Schulp, and P. H. Verburg, Mapping landscape potential for outdoor recreation using different archetypical recreation user groups in the European Union, *Ecological Indicators*, *85*, 105–116, doi: 10.1016/j.ecolind.2017.10.015, 2018.

Korndorfer, J., Dark Skies Cumbria, https://www.friendsofthelakedistrict.org.uk/dark-skies-cumbria, [Online; accessed 28-August-2019], 2019.

Krause, B., Anatomy of the Soundscape, *Journal of the Audio Engineering Society*, *56*(1/2), 2008.

Landis, J. R., and G. G. Koch, The Measurement of Observer Agreement for Categorical Data, *Biometrics*, *33*(1), 159–174, doi: 10.2307/2529310, 1977.

Lansdall-Welfare, T., et al., Content analysis of 150 years of British periodicals, *Proceedings of the National Academy of Sciences of the United States of America*, *114*(4), E457–E465, doi: 10.1073/pnas.1606380114, 2017.

Le, Q., and T. Mikolov, Distributed representations of sentences and documents, in *31st International Conference on Machine Learning, ICML 2014*, vol. 4, pp. 2931–2939, Beijing, China, 2014.

Leidner, J. L., and M. D. Lieberman, Detecting geographical references in the form of place names and associated spatial natural language, in *SIGSPATIAL Special*, vol. 3, pp. 5–11, New York, NY, USA, doi: 10.1145/2047296.2047298, 2011.

Leidner, J. L., G. Sinclair, and B. Webber, Grounding spatial named entities for information extraction and question answering, in *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, vol. 1, pp. 31–38, Stroudsburg, PA, USA, doi: 10.3115/1119394.1119399, 2003.

Lesk, M., Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26, ACM, New York, NY, USA, doi: 10.1145/318723.318728, 1986.

Levin, B., *English Verb Classes and Alternations*, 366 pp., University of Chicago Press, 1993.

Lim, K. H., K. E. Lee, D. Kendal, L. Rashidi, E. Naghizade, S. Winter, and M. Vasardani, The grass is greener on the other side: Understanding the effects of green spaces on Twitter user sentiments, in *WWW'18 Companion: The 2018 Web Conference Companion*, pp. 275–282, Lyon, France, doi: 10.1145/3184558.3186337, 2018.

Liu, B., *Sentiment Analysis and Opinion Mining*, May, 1–108 pp., Morgan & Claypool publishers, Toronto, doi: 10.2200/S00416ED1V01Y201204HLT016, 2012.

Lothian, A., Landscape and the philosophy of aesthetics: Is landscape quality inherent in the landscape or in the eye of the beholder?, *Landscape and Urban Planning*, *44*(4), 177–198, doi: 10.1016/S0169-2046(99)00019-5, 1999.

Lu, Y., M. Castellanos, U. Dayal, and C. Zhai, Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach, in *WWW 2011 – Session: Semantic Analysis*, pp. 347–356, Hyderabad, India, 2011.

Lynott, D., and L. Connell, Modality exclusivity norms for 423 object properties, *Behavior Research Methods*, *41*(2), 558–564, doi: 10.3758/BRM.41.2.558, 2009.

Macfarlane, R., *The Wild Places*, Granta Books, London, 2008.

MacFarlane, R., C. Haggett, D. Fuller, H. Dunsford, and B. Carlisle, Tranquillity Mapping: Developing a Robust Methodology for Planning Support, *Tech. Rep. Report to the Campaign to Protect Rural England, Countryside Agency, North East Assembly, Northumberland Strategic Partnership, Northumberland National Park Authority and Durham County Council, Centre for Environmental & Spatial Analysis, Northumbria Univ*, Northumbria University, Newcastle University, 2004.

Majid, A., and N. Burenhult, Odors are expressible in language, as long as you speak the right language, *Cognition*, *130*(2), 266–270, doi: 10.1016/j.cognition.2013.11.004, 2014.

Majid, A., and S. C. Levinson, The Senses in Language and Culture, *The Senses and Society*, *6*(1), 5–18, doi: 10.2752/174589311X12893982233551, 2011.

Manning, C. D., and H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, doi: 10.1145/601858.601867, 1999.

Mark, D. M., and A. G. Turk, Ethnophysiography, *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pp. 1–11, doi: 10.1002/9781118786352.wbieg0349, 2017.

Marlow, C., M. Naaman, d. boyd, and M. Davis, HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read, in *HT'06 17th Conference on Hypertext and Hypermedia*, pp. 31–39, Odense, Denmark, 2006.

Martins, B., H. Manguinhas, and J. Borbinha, Extracting and exploring the geo-temporal semantics of textual resources, in *Proceedings - IEEE International Conference on Semantic Computing 2008, ICSC 2008*, pp. 1–9, Santa Clara, California, doi: 10.1109/ICSC.2008.86, 2008.

Matasovic, R., *Etymological Dictionary of Proto-Celtic*, BRILL, Leiden, Boston, 2009.

McGregor, S., and B. McGillivray, A Distributional Semantic Methodology for Enhanced Search in Historical Records: A Case Study on Smell, in *14th Conference on Natural Language Processing (KONVENS 2018)*, Vienna, Austria, doi: 10.5281/zen-odo.1403213, 2018.

McLean, K., Smellmap: Edinburgh, https://sensorymaps.com/portfolio/smell-map-edinburgh/, [Online; accessed 16-August-2019], 2011.

Michel, J.-B., et al., Quantitative Analysis of Culture Using Millions of Digitized Books, *Science*, *331*(6014), 176–182, doi: 10.1126/science.1199644, 2011.

Michelin, Y., T. Joliveau, and C. Planchat-Héry, Landscape in Participatory Processes: Tools for Stimulating Debate on Landscape Issues? A Conceptual and Methodological Reflection from Research-Action Projects in France, in *The European Landscape Convention. Challenges of Participation*, edited by M. Jones and M. Stenseke, pp. 145–173, Springer, doi: 10.1007/978-90-481-9932-7_11, 2011.

Millennium Ecosystem Assessment, *Ecosystems and human well-being*, vol. 5, 1–100 pp., doi: 10.1196/annals.1439.003, 2005.

Miller, N. P., US National Parks and management of park soundscapes: A review, *Applied Acoustics*, *69*(2), 77–92, doi: 10.1016/j.apacoust.2007.04.008, 2008.

Moncla, L., W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio, Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus, in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Dallas/Fort Worth, TX, USA, 2014.

Monteiro, B., C. Davis, and F. Fonseca, A survey on the geographic scope of textual documents, *Computers and Geosciences*, *96*, 23–34, doi: 10.1016/j.cageo.2016.07.017, 2016.

Morton, T., *Ecology without nature: rethinking environmental aesthetics*, 264 pp., Harvard University Press, Harvard, 2009.

Mullet, T. C., S. H. Gage, J. M. Morton, and F. Huettmann, Temporal and spatial variation of a winter soundscape in south-central Alaska, *Landscape Ecology*, *31*(5), 1117–1137, doi: 10.1007/s10980-015-0323-0, 2016.

Murrieta-Flores, P., A. Baron, I. Gregory, A. Hardie, and P. Rayson, Automatically Analyzing Large Texts in a GIS Environment: The Registrar General's Reports and Cholera in the 19th Century, *Transactions in GIS*, *19*(2), 296–320, doi: 10.1111/tgis.12106, 2015.

Murrieta-Flores, P., C. Donaldson, and I. Gregory, GIS and Literary History: Advancing Digital Humanities Research through the Spatial Analysis of Historical Travel Writing and Topographical Literature, *Digital Humanities Quarterly*, *11*, 2017.

Nasar, J. L., *Environmental Aesthetics: Theory, research and Application*, 529 pp., Cambridge, 1992.

National Parks UK, The UK's 15 National Parks, https://nationalparks.uk/, [Online; accessed 27-August-2019], 2019.

National Trust, Beatrix Potter, the Lake District and the National Trust, https://www.nationaltrust.org.uk/beatrix-potter-gallery-and-hawkshead/features/beatrix-potter-the-lake-district-and-the-national-trust, [Online; accessed 27-August-2019], 2019a.

National Trust, Our response to the use of 4x4s and motorbikes on the unclassified road between Tilberthwaite Farm and Little Langdale, https://www.nationaltrust.org.uk/features/our-response-to-the-use-of-4x4s-and-motorbikes-on-the-unclassified-road-between-tilberthwaite-farm-and-little-langdale, [Online; accessed 28-August-2019], 2019b.

Natural England, Areas of outstanding natural beauty (AONBs): designation and management, https://www.gov.uk/guidance/areas-of-outstanding-natural-beauty-aonbs-designation-and-management, [Online; accessed 14-August-2019], 2019.

Naveh, Z., and A. Lieberman, *Landscape Ecology: theory and application*, 2nd ed., 360 pp., Springer-Verlag, New York, 1994.

Neis, P., M. Goetz, and A. Zipf, Towards automatic vandalism detection in open street map, *ISPRS International Journal of Geo-Information*, *1*(3), 315–332, doi: 10.3390/ijgi1030315, 2012.

Nicholson, B., Counting culture; Or, how to read Victorian newspapers from a distance, *Journal of Victorian Culture*, *17*(2), 238–246, doi: 10.1080/13555502.2012.683331, 2012.

Nielsen, J., The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities, http://www.useit.com/alertbox/participation_inequality.html, [Online; accessed 23-August-2019], 2006.

Nomination, Lake District Nomination, http://www.lakedistrict.gov.uk/caringfor/projects/whs/lake-district-nomination, [Online; accessed 27-August-2019], 2017.

Olteanu, A., C. Castillo, F. Diaz, and E. Kiciman, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, *Frontiers in Big Data*, *2*(13), doi: 10.3389/fdata.2019.00013, 2019.

Olwig, K. R., Recovering the substantive nature of landscape, *Annals of the Association of American Geographers*, *86*(4), 630–653, doi: 10.1111/j.1467-8306.1996.tb01770.x, 1996.

Olwig, K. R., *Landscape, nature, and the body politic : from Britain's renaissance to America's new world*, 299 pp., University of Wisconsin Press, Madison, 2002.

Olwig, K. R., Representation and alienation in the political land-scape, *Cultural Geographies*, *12*, 19–40, doi: 10.1191/1474474005eu3210a, 2005.

Overell, S., and S. Rüger, Using co-occurrence models for placename disambiguation, *International Journal of Geographical Information Science*, *22*(3), 265–287, doi: 10.1080/13658810701626236, 2008.

Palmer, C., and E. Brady, Landscape and value in the work of Alfred Wainwright (1907-1991), *Landscape Research*, *32*(4), 397–421, doi: 10.1080/01426390701449778, 2007.

Palmer, J. F., Using spatial metrics to predict scenic perception in a changing landscape: Dennis, Massachusetts, *Landscape and Urban Planning*, *69*(2-3), 201–218, doi: 10.1016/j.landurbplan.2003.08.010, 2004.

Pang, B., L. Lee, and S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, *Empirical Methods in Natural Language Processing (EMNLP)*, *10*(July), 79–86, doi: 10.3115/1118693.1118704, 2002.

Parveen, N., Ban 4x4 off-roading in Lake District, campaigners say, https://www.theguardian.com/uk-news/2018/sep/13/ban-4x4-off-roading-in-the-lake-district-campaigners-say, [Online; accessed 28-August-2019], 2018.

Pavord, A., *Landskipping*, Bloomsbury Publishing Plc, 2017.

Pennington, J., R. Socher, and C. D. Manning, GloVe: Global Vectors for Word Representation, in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, doi: 10.3115/v1/D14-1162, 2014.

Pérez-Martínez, G., A. J. Torija, and D. P. Ruiz, Soundscape assessment of a monumental place: A methodology based on the perception of dominant sounds, *Landscape and Urban Planning*, *169*, 12–21, doi: 10.1016/j.landurbplan.2017.07.022, 2018.

Pheasant, R., K. Horoshenkov, G. Watts, and B. Barrett, The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places?, *The Journal of the Acoustical Society of America*, *123*(3), 1446–1457, doi: 10.1121/1.2831735, 2008.

Pheasant, R. J., and G. R. Watts, Towards predicting wildness in the United Kingdom, *Landscape and Urban Planning*, *133*, 87–97, doi: 10.1016/j.landurbplan.2014.09.009, 2015.

Pidd, H., National Trust buys Lake District hill revered by Turner for its views, https://www.theguardian.com/uk-news/2019/jun/05/national-trust-adds-lake-district-vista-painted-by-turner-to-portfolio, [Online; accessed 27-August-2019], 2019.

Pieretti, N., A. Farina, and D. Morri, A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI), *Ecological Indicators*, *11*(3), 868–873, doi: 10.1016/j.ecolind.2010.11.005, 2011.

Pijanowski, B. C., A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, What is soundscape ecology? An introduction and overview of an emerging new science, *Landscape Ecology*, *26*(9), 1213–1232, doi: 10.1007/s10980-011-9600-8, 2011a.

Pijanowski, B. C., L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, Soundscape Ecology: The Science of Sound in the Landscape, *BioScience*, *61*(3), 203–216, doi: 10.1525/bio.2011.61.3.6, 2011b.

Plieninger, T., S. Dijks, E. Oteros-Rozas, and C. Bieling, Assessing, mapping, and quantifying cultural ecosystem services at community level, *Land Use Policy*, *33*, 118–129, doi: 10.1016/j.landusepol.2012.12.013, 2013.

Prior, J., Sonic environmental aesthetics and landscape research, *Landscape Research*, *42*(1), 6–17, doi: 10.1080/01426397.2016.1243235, 2017.

Purves, R. S., and C. Derungs, From Space to Place: Place-Based Explorations of Text, *International Journal of Humanities and Arts Computing*, *9*(1), 74–94, doi: 10.3366/ijhac.2015.0139, 2015.

Purves, R. S., A. J. Edwardes, and J. Wood, Describing place through user generated content, *First Monday. Peer-reviewed journal on the internet*, *16*(9), 2011.

Purves, R. S., P. Clough, C. B. Jones, M. H. Hall, and V. Murdock, *Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*, Now Foundations and Trends, 2018.

Pustejovsky, J., and A. Stubbs, *Natural Language Annotation for Machine Learning*, O'reilly, Sebastopol, CA, 2012.

Quercia, D., and R. Schifanella, Smelly Maps: The Digital Life of Urban Smellscapes, in *9th International AAAI Conference on Web and Social Media*, Oxford, UK, 2015.

Rattenbury, T., and M. Naaman, Methods for extracting place semantics from Flickr tags, *ACM Transactions on the Web*, *3*(1), 1–30, doi: 10.1145/1462148.1462149, 2009.

Regier, T., A. Carstensen, and C. Kemp, Languages Support Efficient Communication about the Environment: Words for Snow Revisited, *PLoS ONE*, *11*(4), 1–17, doi: 10.1371/journal.pone.0151138, 2016.

Relph, E., *Place and Placelessness*, Pion Press, London, 1976.

Resch, B., A. Summa, P. Zeile, and M. Strube, Citizen-centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm, *Urban Planning*, *1*(2), 114–127, doi: 10.17645/up.v1i2.617, 2016.

Richards, D. R., and D. A. Friess, A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs, *Ecological Indicators*, *53*, 187–195, doi: 10.1016/j.ecolind.2015.01.034, 2015.

Richards, D. R., and B. Tunçer, Using image recognition to automate assessment of cultural ecosystem services from social media photographs, *Ecosystem Services*, *31*, 318–325, doi: 10.1016/j.ecoser.2017.09.004, 2018.

Rose, T., M. Stevenson, and M. Whitehead, The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources, in *Proceedings of the Third International Conference on Language Resources and Evaluation (2002)*, vol. 1, pp. 827–833, Las Palmas, Canary Islands, Spain, 2002.

Rykiel, E. J., Testing ecological models: The meaning of validation, *Ecological Modelling*, *90*(3), 229–244, doi: 10.1016/0304-3800(95)00152-2, 1996.

San Roque, L., et al., Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies, *Cognitive Linguistics*, *26*(1), 31–60, doi: 10.1515/cog-2014-0089, 2015.

Schafer, R. M., *The soundscape: Our sonic environment and the tuning of the world*, 320 pp., Inner Traditions/Bear & Co, 1993.

Schirpke, U., E. Tasser, and U. Tappeiner, Predicting scenic beauty of mountain regions, *Landscape and Urban Planning*, *111*(1), 1–12, doi: 10.1016/j.landurbplan.2012.11.010, 2013.

Scottish Natural Heritage, National Scenic Areas, https://www.nature.scot/professional-advice/safeguarding-protected-areas-and-species/

protected-areas/national-designations/national-scenic-areas, [Online; accessed 14-August-2019], 2019.

Selman, P., and C. Swanwick, On the meaning of natural beauty in landscape legislation, *Landscape Research*, *35*(1), 3–26, doi: 10.1080/01426390903407160, 2010.

Seresinhe, C., T. Preis, and H. Moat, Quantifying the Impact of Scenic Environments on Health, *Scientific Reports*, *5*(16899), 1–9, doi: 10.1038/srep16899, 2015.

Seresinhe, C., T. Preis, and M. Susannah, Using deep learning to quantify the beauty of outdoor places, *Royal Society Open Science*, *4*(7), doi: 10.1098/rsos.170170, 2017.

Shepherd, N., *The Living Mountain*, Canongate Books, London, 2011.

Smith, D., and G. Crane, Disambiguating Geographic Names in a Historical Digital Library, *5th European Conference on Digital Libraries*, *2163*, 127–136, doi: 10.1007/3-540-44796-2_12, 2001.

Smith, S. J., Soundscape, *Area*, *26*, 232–240, 1994.

Southworth, M., The Sonic Environment of Cities, *Environment and Behavior*, *June*, 49–70, 1969.

Stadler, B., R. Purves, and M. Tomko, Exploring the Relationship Between Land Cover and Subjective Evaluation of Scenic Beauty through User Generated Content, in *Proceedings of the 25th International Cartographic Conference*, International Cartographic Association, Paris, 2011.

Swanwick, C., and G. Fairclough, Landscape Character: Experience from Britain, in *Routledge Handbook of Landscape Character Assessment*, edited by G. Fairclough, I. Sarlöv Herlin, and C. Swanwick, pp. 21–36, Routledge, 2018.

Taylor, J., Echoes in the Mountains: The Romantic Lake District's Soundscape, *Studies in Romanticism*, *57*(3), 2018.

Taylor, J., I. Gregory, and C. Donaldson, Combining Close and Distant Reading: A Multiscalar Analysis of the English Lake District's Historical Soundscape, *International Journal of Humanities and Arts Computing*, *12*, 163–182, 2018.

Taylor, J. G., K. J. Czarnowski, and S. Flick, The importance of water to Rocky Mountain National Park visitors: an adaptation of visitor-employed photography to natural resources management, *Journal of Applied Recreation Research*, *20*(1), 61–85, 1995.

Tenerelli, P., U. Demšar, and S. Luque, Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes, *Ecological Indicators*, *64*, 237–248, doi: 10.1016/j.ecolind.2015.12.042, 2016.

Thoreau, H. D., *Walden*, 370 pp., Yale University Press, New Haven, 1854.

Tieskens, K. F., B. T. Van Zanten, C. J. Schulp, and P. H. Verburg, Aesthetic appreciation of the cultural landscape through social media: An analysis of revealed preference in the Dutch river landscape, *Landscape and Urban Planning*, *177*(June 2017), 128–137, doi: 10.1016/j.landurbplan.2018.05.002, 2018.

Toivonen, T., V. Heikinheimo, C. Fink, A. Hausmann, T. Hiippala, O. Järv, H. Tenkanen, and E. Di Minin, Social media data for conservation science: A methodological overview, *Biological Conservation*, *233*, 298–315, doi: 10.1016/j.biocon.2019.01.023, 2019.

Truax, B., *Handbook for Acoustic Ecology*, ARC Publications, Burnaby, B.C. Canada, 1978.

Tudor, C., An Approach to Landscape Character Assessment, *Tech. Rep. October*, Natural England, 2014.

UK Government, Equality Act 2010, https://www.gov.uk/guidance/equality-act-2010-guidance, [Online; accessed 13-September-2019], 2013.

van Zanten, B. T., D. B. van Berkel, R. K. Meetemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg, Continental scale quatification of landscape values using social media data, *Proceedings of the National Academy of Sciences*, pp. 1–7, doi: 10.1073/pnas.1614158113, 2016.

Vasilescu, F., P. Langlais, and G. Lapalme, Evaluating Variants of the Lesk Approach for Disambiguating Words, in *Conference on Language Resources and Evaluation*, pp. 633–636, Lisbon, Portugal, 2004.

Volk, M., N. Bubenhofer, A. Althaus, and M. Bangerter, Classifying named entities in an alpine heritage corpus, *KI*, *23*, 40–43, 2009.

Wainwright, A. (Ed.), *A Pictorial Guide to the Lakeland Fells. The Northern Fells.*, vol. 5, Frances Lincoln, 1962.

Wainwright, A. (Ed.), *A Pictorial Guide to the Lakeland Fells. The Weastern Fells.*, vol. 7, Frances Lincoln, 1966.

Wallgrün, J. O., M. Karimzadeh, A. M. MacEachren, and S. Pezanowski, GeoCorpora: building a corpus to test and train microblog geoparsers, *International Journal of Geographical Information Science*, *32*(1), 1–29, doi: 10.1080/13658816.2017.1368523, 2018.

Wartmann, F. M., E. Egorova, C. Derungs, D. M. Mark, and R. S. Purves, More Than a List: What Outdoor Free Listings of Landscape Categories Reveal about Commonsense Geographic Concepts and Memory Search Strategies, in *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*, edited by S. I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, and S. Bell, pp. 224–243, Springer International Publishing, doi: 10.1007/978-3-319-23374-1_11, 2015.

Wartmann, F. M., E. Acheson, and R. S. Purves, Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach, *International Journal of Geographical Information Science*, *32*(8), 1–21, doi: 10.1080/13658816.2018.1445257, 2018.

Watkins, D., Lake District National Park. Landscape Character Assessment and Guidelines, *Tech. rep.*, Chris Blandford Associates environment landscape planning, 2008.

Webb, E. J., D. T. Campbell, R. D. Schwartz, and L. Sechrest, *Unobtrusive measures: Nonreactive research in the social sciences*, Rand Mcnally, Oxford, England, 1966.

Williams, J. M., Synaesthetic Adjectives: A Possible Law of Semantic Change, *Language*, *52*(2), 461, doi: 10.2307/412571, 1976.

Winter, B., M. Perlman, and A. Majid, Vision dominates in perceptual language: English sensory vocabulary is optimized for usage, *Cognition*, *179*(May), 213–220, doi: 10.1016/j.cognition.2018.05.008, 2018.

Wohl, A. S., Octavia Hill and The Homes of the London Poor, *Journal of British Studies*, *10*(2), 105–131, doi: 10.1086/385612, 1971.

Wordsworth, D., *Journals of Dorothy Wordsworth*, vol. 1, Macmillan and Co., New York, 1897.

Wordsworth, W., *A Complete Guide to the Lakes: Comprising Minute Directions for the Tourist, with Mr. Wordsworth's Description of the Scenery of the Country, &c. and Three Letters on the Geology of the Lake District, by the Late Professor Sedgwick.*, J. Hudson, 1843.

Xu, S., A. Klippel, A. M. MacEachren, and P. Mitra, Exploring Regional Variation in Spatial Language Using Spatially Stratified Web-Sampled Route Direction Documents, *Spatial Cognition and Computation*, *14*(4), 255–283, doi: 10.1080/13875868.2014.943904, 2014.

Yang, B.-E., and T. J. Brown, A Cross-Cultural Comparison of Preferences for Landscape Styles and Landscape Elements, *Environment and Behavior*, *24*(4), 471–507, 1992.

Zimmer, M., Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity, *Social Media and Society*, *4*(2), doi: 10.1177/2056305118768300, 2018.

A

APPENDIX: PUBLICATION 1

Chesnokova, O., Nowak, M., and Purves, R.S., 2017. A crowdsourced model of landscape preference. In: E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore, eds. 13th International Conference on Spatial Information Theory (COSIT 2017). Leibniz International Proceedings in Informatics, 19:1-19:13.

PhD candidate's contributions: Co-developing research idea based on the same datasets (ScenicOrNot and Geograph UK) as the master thesis of Mario Nowak, data processing and analysis, creation of maps, writing the draft manuscript and incorporating co-author's feedback.

A Crowdsourced Model of Landscape Preference

Olga Chesnokova[1], Mario Nowak[2], and Ross S. Purves[3]

1 **Department of Geography, University of Zürich, Zürich, Switzerland**
`olga.chesnokova@geo.uzh.ch`
2 **Department of Geography, University of Zürich, Zürich, Switzerland**
`mario.h.nowak@gmail.com`
3 **Department of Geography, University of Zürich, Zürich, Switzerland**
`ross.purves@geo.uzh.ch`

Abstract

The advent of new sources of spatial data and associated information (e.g. Volunteered Geographic Information (VGI)) allows us to explore non-expert conceptualisations of space, where the number of participants and spatial extent coverage encompassed can be much greater than is available through traditional empirical approaches. In this paper we explore such data through the prism of landscape preference or *scenicness*. VGI in the form of photographs is particularly suited to this task, and the volume of images has been suggested as a simple proxy for landscape preference. We propose another approach, which models landscape aesthetics based on the descriptions of some 220000 images collected in a large VGI project in the UK, and more than 1.5 million votes related to the perceived scenicness of these images collected in a crowdsourcing project. We use image descriptions to build features for a supervised machine learning algorithm. Features include the most frequent uni- and bigrams, adjectives, presence of verbs of perception and adjectives from the "Landscape Adjective Checklist". Our results include not only qualitative information relating terms to scenicness in the UK, but a model based on our features which can predict some 52% of the variation in scenicness, comparable to typical models using more traditional approaches. The most useful features are the 800 most frequent unigrams, presence of adjectives from the "Landscape Adjective Checklist" and a spatial weighting term.

1998 ACM Subject Classification I.7.0 Document and Text Processing

Keywords and phrases VGI, crowdsourcing, semantics, landscape preference

Digital Object Identifier 10.4230/LIPIcs.COSIT.2017.19

1 Introduction

The advent of new sources of spatial data, and in particular those which are generated not through a top-down, regulated process, but bottom-up, by individuals with varying backgrounds and motivations, has brought with it new opportunities for research. In particular, the advent of spatial data associated with natural language, typically in the form of tags or unstructured text provide a potential route to exploring ways in which space is described in language, albeit typically in corpora where we as researchers have very little control. The data studied in such research can be produced in a number of ways, and differing, but overlapping, definitions have been assigned to such data including those related to volunteered geographic information (VGI), crowdsourcing, user-generated content, social media, citizen science and so on [7]. These definitions are important since they have implications for the ways in which data are produced, and in turn the ways in which they can reasonably be interpreted.

13th International Conference on Spatial Information Theory (COSIT 2017).
Editors: Eliseo Clementini, Maureen Donnelly, May Yuan, Christian Kray, Paolo Fogliaroni, and Andrea Ballatore;
Article No. 19; pp. 19:1–19:13

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

One obvious, and much studied, source of such data are the tags and descriptions associated with georeferenced images. Here, researchers typically assume that images and their descriptions often capture information about named locations, their properties and, occasionally, notions related to sense of place (e.g. [28, 16]). Indeed, Fisher and Unwin [8] presciently recognised this potential in 2005, stating that "GI theory articulates the idea of absolute Euclidean spaces quite well, but the socially-produced and continuously changing notion of place has to date proved elusive to digital description except, perhaps, through photography and film. (p. 6)." Nonetheless, in practice analysing text and extracting information related to place has proved challenging, and many studies have either focussed above all on exploring the properties of text related to location, with limited or no opportunities for validation, or on using counts of images as a proxy for some spatially varying phenomena and generating appropriate statistical models (e.g. [5, 36, 37]).

The act of georeferencing images typically implies that an individual wishes to relate a particular image to an event (not relevant in the context of this work) or a location. The act of producing an image however is not random, and neither is the act of choosing to share an image with others in an online source [11]. Images capturing locations presumably capture perceptually salient elements of a landscape, and thus, accompanied by their descriptions might provide us with clues as to how landscape is conceptualised and parcelled up into cognitive entities [22]. Understanding landscape, and the ways in which it is perceived is not merely an abstract research question, but one with considerable direct policy and societal relevance, since landscapes are the subject of national and international policies and regulation. Contemporaneously with the emergence of new data sources such as those described above, has been an increasing realisation in many areas of policy that there is a need to include not only top-down definitions of landscapes in policy work, but also to capture bottom-up ways in which landscapes are perceived and experienced. Even seemingly simple notions such as landscape aesthetics have proved remarkably challenging to generalise and model spatially, and although methods based in the social sciences can capture well the diversity of opinions about individual locations, they are ill-suited to characterising large regions [37].

In this paper we set out to demonstrate, through the use of two, related, datasets, how we can firstly, capture through textual descriptions, elements of a landscape which are perceived as more or less attractive across a large region. To do so, we combine descriptions of georeferenced images which are an excellent example of VGI *sensu* Goodchild [12] with a large crowdsourced data containing *scenicness* rating for more than 220000 images. We then develop and evaluate a predictive model of scenicness, which as its primary input uses text describing images, and thus aims to model scenicness as a function of language.

1.1 Related work

In the following we briefly set out related work from two key areas. Firstly, we summarise concepts related to landscape aesthetics and its assessment. Secondly, we explore examples of research which have used novel data sources to explore landscape properties in a range of ways.

Theories seeking to explain landscape perception and aesthetics typically focus on both evolutionary and cultural influences [19, 15]. Evolutionary approaches assume that preferences with respect to landscape relate to the ability of landscapes to meet human needs such as 'prospect' (i.e. the ability to command a landscape through sight) and 'refuge' (the potential to conceal oneself in a landscape) [1]. Other, related concepts include the ability to 'make sense' of the environment (coherence and legibility of landscapes), and 'involvement' or ability

to function well in the environment (complexity and mystery of landscapes) [18]. Cultural influences on landscape preference are recognised in the emergence of work on landscape and language, for example, through the study of ethnophysiography [22] which notes the importance of cultural influences and the absence of universally shared landscape elements.

Irrespective of the theoretical perspective taken, typical approaches to capturing landscape perception have focussed on in-situ methods using, for example, interviews and participatory mapping [2, 27]. However, the need to be on site makes such approaches poorly suited to capturing dynamic landscape preferences over large areas, and also makes it difficult to control potential influences. Such limitations, and the simple need to generate more lab-based reproducible experiments, led to the development of approaches based around photographs of landscapes where participants can be presented with images controlling the visual field [31], seasonal changes, or introducing extra factors (e.g. presence of animals [17] or anthropogenic objects [20]).

The advent of VGI, and the realisation that such data might contain diverse, independent and decentralised information, provided opportunities to replicate previous work on geographic concepts [34], and to demonstrate that such data were a reliable source of information about landscape characteristics and the ways in which landscapes were categorised [6, 28]. In parallel, the need to generate landscape indicators related to cultural ecosystem services and landscape preferences over large areas has led some of researchers to use the position and number of images taken as a proxy indicator of landscape preference [36, 37], or to incorporate the number of individuals taking pictures [3, 11] and their origins [10]. Others have realised that the images themselves contain information central to understanding landscape preference, and have analysed image content to explore cultural ecosystem services [30]. The importance of scenicness in a policy context, and the possibilities offered by new data sources are recognised in recent work exploring the link between wellbeing and scenicness using crowdsourced data, and attempting to model scenicness using user generated content [33, 32].

In this paper we seek to build on previous work in two key ways. Firstly, in-situ and lab-based studies of landscape preference have typically worked, of necessity, with relatively small groups of participants in focussed, often coherent, landscapes. Our study, by using VGI at the scale of Great Britain, allows us to explore landscape preferences across a whole country, and to explore regional differences between such preferences. Secondly, attempts to model scenicness have typically focussed on using spatial data in some form as explanatory variables (for example number of images, elevation, number of visible pixels, landcover type, etc.). We take an approach which we argue is likely to be closer to the way in which a particular landscape is perceived, and build a model of scenicness which uses language (in the form of words and phrases extracted from written descriptions) as explanatory variables.

In the following, we first describe the datasets on which we carried out our experiment, and the steps we took in processing, analysing and modelling scenicness with these data. We then present our results, demonstrating that the words used to describe scenic areas make clear distinctions especially between scenes perceived to be more or less anthropogenically influenced. Our model of scenicness is capable of explaining about 52% of the variance in scenicness in space, which is comparable to typical state of the art approaches. We then discuss the implications of these results, before concluding with some suggestions for future research.

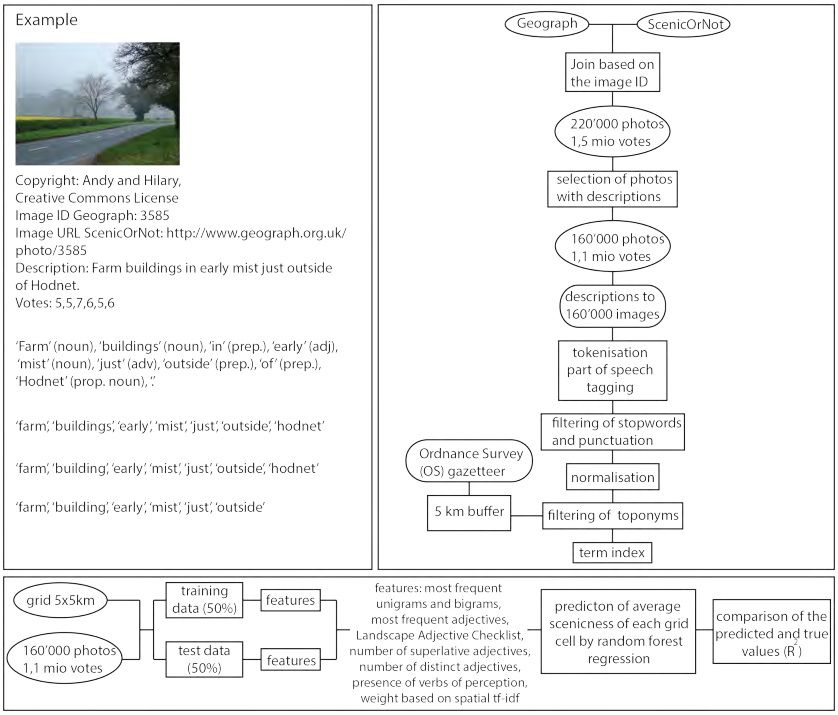


Figure 1 Steps of the data acquisition and preprocessing with an example.

2 Data and methods

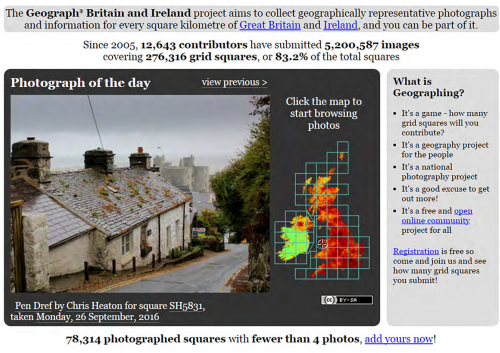
As set out above, our aims are twofold. Firstly, we wish to identify which terms are typically used with more or less scenic images, as described by votes in ScenicOrNot project and, secondly, based only on terms describing images to develop a spatially contiguous model of scenicness at the country level. In the following we describe the datasets used, and in particular aspects relevant to our work. We then set out our approach to processing the corpus, before describing the features used in producing our spatial model of scenicness. Fig. 1 gives a visual overview of the material which follows.

2.1 Data and study region

We use two unique, and related, datasets in this work. The Geograph[1] dataset (Fig. 2a) is a crowdsourced collection with more than 12000 contributors, launched in 2005, with the aim of collecting "geographically representative photographs and information for every square kilometre of Great Britain and Ireland." The project takes the form of a game, with users receiving points for uploading georeferenced images and associated descriptions, and content is moderated. The entire dataset is available under a Creative Commons Licence, and in this paper we used a version downloaded in June 2016 consisting of ca. five million images.

The ScenicOrNot[2] project (Fig. 2b) was initiated in 2009 by MySociety and is currently hosted by the Data Science Lab at Warwick Business School. The goal of the project is

[1] http://www.geograph.org.uk/
[2] http://scenicornot.datasciencelab.co.uk/



(a) Interface of the Geograph project

(b) Interface of the ScenicOrNot project

Figure 2 Interface of the Geograph project (Copyright Chris Heaton, Creative Commons Licence) and the ScenicOrNot project, where users can rate a Geograph photograph from 1 (not scenic) to 10 (very scenic).

to crowdsource scenicness ratings using Geograph images. In contrast to Geograph, where it is reasonable to assume that users uploading images typically also took the pictures in question (and thus visited the landscape), the ScenicOrNot project is purely internet based. Participants, about whom no demographic information is collected, are presented with a series of random images, with neither associated locations or descriptions, and asked to rate them on a scale of 1 (not scenic) to 10 (scenic) for scenicness. More than 220000 Geograph images had amassed some 1.5 million votes by June 2016 in the ScenicOrNot collection.

In the following our corpus consists of the 160000 Geograph images which both have a description, and are associated with three or more votes in ScenicOrNot.

2.2 Corpus processing

Our aim in corpus processing was to explore how terms used in describing Geograph images were associated with scenicness ratings. Since our starting point are natural language captions, standard corpus processing steps were applied. In the following, we briefly describe these steps, which were, in the main, carried out using the Python-based NLTK[3] library.

Each image description was in parallel tokenised, and part of speech tagged. The tokens were then filtered for stopwords and punctuation, before being normalised by changing all tokens to lower case and reducing tokens to their lemmas. Our aim was to build a term index, with associated features, for use in exploring the semantics of scenic locations.

Since we were explicitly not interested in the names of locations, we filtered toponyms from descriptions using gazetteer look-up in a 5km window around the coordinates associated with images. We used a freely available gazetteer, based on the 1:50000 maps from the Ordnance Survey for this process. This approach aims to strike a balance between removing local toponyms, which may be the subject of considerable semantic ambiguity (e.g. does bath refer to a place to bathe or the historic city) and retaining tokens which are being used in a non-toponymic sense.

Having performed these steps we are left with a term index, where unique entries are made up of tuples containing normalised tokens (unigrams and bigrams) present once or more in a description, part of speech tagging and the images IDs with which they are associated. Since

[3] `http://www.nltk.org/`

each term can be present in one or more images, and each image is ranked three or more times, we assign an average scenicness to every term in our index. Importantly, identical tokens having different parts of speech will have different values of average scenicness. Furthermore, since we store image IDs, we also have access to all locations associated with a term, the array of votes and an overall frequency of the term, based on the number of images described using a given term. Using our term index, it is possible to generate lists of terms, ranking or filtering by, for example, average scenicness, part of speech or frequency.

2.3 Feature choice and modelling scenicness

The final step in our approach was to create a spatially contiguous model of scenicness based on our term index. We predict scenicness for 5km grid cells, using Random Forests regression, which is a state of the art non-linear, non-parametric method in supervised machine learning, and which requires no assumptions with respect to the data distribution [4]. Our choice of 5km was motivated by the underlying 1km granularity of the Geograph data and its associated spatial distribution. We report briefly on sensitivity to resolution in the discussion.

A key task in creating such a model is the choice of appropriate features. Our basic approach was to use training data associated with 5km grid cells, where average scenicness was associated with features based on our terms. Only descriptions consisting of at least five tokens, after filtering as described above, were used in the model. The simplest possible feature set would be one based purely on unigrams, that is to say individual tokens from image descriptions found in grid cells (e.g. 'hill', 'mountain', 'shop', etc.).

However, in natural language processing [21] it is typical to also consider n-grams, and here we also experimented with bigrams (e.g. sequences of two tokens such as 'steep hill', 'rugged mountain', 'closed shop') as features. By reducing the feature space it is often possible to maintain model predictive capacity, while improving performance, and we also experimented by reducing the number of unigrams considered to the n-most frequent. Other features of our data, and previous work on landscape description, suggest additional potential model features which are listed below:

- adjectives alone: since adjectives are assumed to be strong indicators of subjectivity and sentiment; [14], we used unigrams consisting only of frequent adjectives;
- "Landscape Adjective Checklist": presence of adjectives pertaining specifically to landscape in Craik's list [24];
- the number of superlative adjectives as identified during part of speech tagging, with the assumption that superlatives are more likely to be used in more scenic areas;
- the number of distinct adjectives found in a description, with the assumption that more adjectives are used in more scenic areas;
- the presence of a verb of perception [39], where we assume that the presence of verbs of perception may indicate descriptions more relevant with respect to scenicness (e.g. by reducing the weight of descriptions focussing on historical events at a location);
- a weight based on spatial tf-idf [29]: here terms which are used frequently in an individual grid cell, but rarely in the collection as a whole are given a higher weight.

2.4 Training and test data

In any supervised model it is necessary to generate both training and test datasets. However, the way in which the data are split can have important implications for not only the quality of the model, but also for any implications which can be drawn from the results. Since an important property of crowdsourced data are user-generated biases in data production [13], we

(a) Scenicness between 1–3, $n_{images} = 14072$, $n_{users} = 2137$.

(b) Scenicness between 3–5, $n_{images} = 155822$, $n_{users} = 4170$.

(c) Scenicness between 5–7, $n_{images} = 79752$, $n_{users} = 3340$.

(d) Scenicness between 7–10, $n_{images} = 3134$, $n_{users} = 851$.

Figure 3 Average scenicness for 150 most frequent nouns extracted from image descriptions - font size indicates relative frequency within scenicness range.

considered these, as well as the desired spatial contiguity of our model in generating training and test data. Thus, our models were trained (and tested) on the following configurations, with 50% assigned to training and test data respectively in both cases:

- fully random: image descriptions are simply selected at random from the full corpus;
- user dependent random: since we expect individual users to write characteristic descriptions, and since Geograph is subject to participation inequality, meaning that a single user may contribute a large proportion of the descriptions in a single area, we select random images while allowing individual users only to appear in either training or test datasets.

3 Results and interpretation

3.1 Semantics of scenicness

The word clouds in Fig. 3 exemplify our results, illustrating the average scenicness of nouns after part of speech tagging of image descriptions. A number of features are worthy of observation here. At a first glance, the lowest rated scenicness values are related to nouns which are clearly in developed areas (e.g. ‘motorway’, ‘housing’, ‘shop’, ‘stadium’). The highest rated scenicness nouns include Gaelic words, terms related to natural processes, wildlife and some esoteric examples (e.g. ‘coire’, ‘avalanche’, ‘otter’, ‘backcloth’). However, these classes contain a small proportion of the total set (ca. 6%), with only some 1% of nouns being found in the most scenic class. Thus, many of these nouns belong to the long tail of our data, and although they reflect a clear split between developed areas and more natural landscapes (associated with Gaelic placenames in the Highlands of Scotland) we should be careful not to overinterpret these terms.

Unsurprisingly, since each image was rated at least three times, and many of the nouns are associated with multiple images, the vast majority (94%) of nouns have average scenicness

ratings of between 3 and 7. Exploring these classes, it becomes apparent that the clear split so visible in the two extreme classes is much less prominent. Thus, we find that nouns such as 'village', 'lane' and 'wood' are all rated on average 3–5, even though these might be terms typically expected to be associated with more rural, and thus potentially more scenic images. However, exploring the nouns rated 5–7 it again becomes clear that differences exist. Here, many more nouns appear to relate to perceived natural (as opposed to rural) scenes (e.g. 'moorland', 'summit', 'ridge').

3.2 Predicting scenicness

We tested the goodness of fit of our Random Forest regression using the features as described above, and two different configurations of test and training data. Independent of the configuration chosen, we only predicted scenicness values for grid cells where at least two descriptions were present in both training and test data.

Goodness of fit improved as we increased the number of unigrams in the model until we reached the 800 most frequent unigrams. Including presence of adjectives from the "Landscape Adjective Checklist" by Craik and weighting according to spatial tf-idf further increased goodness of fit to a maximum value of around 52% (52.4% in the case of fully random and 52.0% in the case of user dependent random division on training and test data).

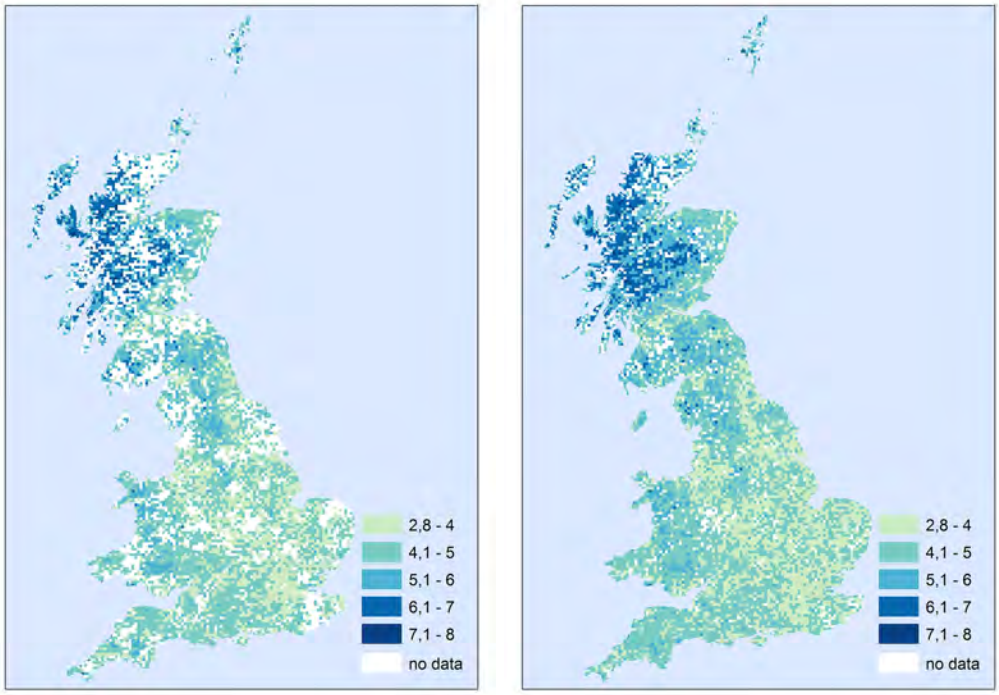
Fig. 4 shows the spatial pattern of predicted scenicness for both configurations. Particularly evident here are the larger number of grid cells for which no value could be predicted where training and test data were randomly selected according to users. Here, the effects of participation inequality result in many grid cells where the majority of images and associated descriptions were taken by a single user, and we thus cannot predict scenicness. However, given the limited variation in model goodness of fit, it appears that this restriction may be unnecessary.

A further important issue in our model is the existence of spatial autocorrelation in model residuals. Testing for Morans-I revealed values of around 0.12 according to model configuration, implying that the chances of random clustering in our model are less than 1%. A typical approach to assessing the influence of spatial autocorrelation in Random Forest regression is therefore to include grid centroids as features in the model [23]. Doing so increased goodness of fit to 56% and reduced spatial autocorrelation in the residuals to 0.05. An alternative model including spatial information by assigning county names (administrative units) to every image, resulted in a decrease of Morans's I to 0.10, with goodness of fit remaining at 52%. This approach includes local neighbourhood relationships and more natural divisions of landscape (since at least in the UK county boundaries typically are a mix of the *fiat* and *bona fide*). Since model results for a model based only on language and containing additional explicit spatial information are similar we thus conclude that our results are not biased by spatial autocorrelation [37].

4 Discussion

In this paper we explored the use of two, related datasets which were both generated by the crowd, though in very different ways, to understand how landscape, and in particular scenicness is captured in language.

Our results were generated after a typical natural language pipeline to tokenise, classify and filter image descriptions. Importantly, we also included a step to remove toponyms from image descriptions, since we were not interested in the names of scenic places, but rather in their properties. Our results demonstrate a clear transition from nouns associated with



(a) User-based division. **(b)** Random division.

Figure 4 Maps of the scenicness prediction results with 'user dependent random division' and 'fully random division'.

urban, developed scenes through more rural landscapes to natural landscapes and a long tail of nouns associated with the Highlands of Scotland. This long tail also reveals one limitation of our approach, since our natural language processing methods cannot deal with Gaelic, and some misclassified words remained in the list of nouns (e.g. *ruadh* refers to the colour red in Gaelic and is commonly used in toponyms i.e. *Sgurr Ruadh* refers to the Red Peak).

Exploration of the word clouds (Fig. 3) reveals that the scenicness of individual terms sometimes contradicts classic ideas in work on landscape preference. For example, water is commonly associated with scenic landscapes [40, 31], yet in our word clouds it has an average scenicness of only 3–5. On closer examination it becomes apparent that water lies in a word cloud containing many rural terms, and the presence of water is common in such scenes. However, at least in our data, rural as opposed to perceived natural scenes are less highly rated. Thus, treating individual nouns (or terms in general) as predictors of scenicness is difficult, and our word clouds reveal more information about the complex interplay between language and landscape. They further indicate the importance of using language, as opposed to purely data-driven approaches to exploring landscape. Approaches extracting landscape properties using intrinsic landscape qualities from standard spatial datasets and associating these with landscape preferences (e.g. [9, 38]) based on ideas of evolutionary-driven landscape perception [18] are unlikely to capture variation of the nature we observe here. Furthermore, our word clouds are potentially powerful tools for generating datasets containing imagery for use in landscape preference experiments and modelling, since they provide an empirical basis for terms used in selecting candidate images, as opposed to approaches based on introspective reasoning or intrinsic, evolutionary determined preferences (cf. [37]) to generate candidate keywords for querying.

Our model of scenicness, irrespective of training data is able to explain some 52% of the variation in scenicness. This is comparable with typical results in more traditional approaches based on interviews or participatory methods [25], approaches using land cover data [35] and work at a continental scale using social media [37]. Although the explained variance is not strongly influenced by our choice of training data, the total number of grid cells for which average scenicness value can be predicted varies by some 20% from around 7000 cells where individual users are only allowed to be present in either test or training data, to 9000 cells where image descriptions are randomly assigned to test or training data. Furthermore, this variation is strongly spatially autocorrelated, with, for example, a single user having taken some 11000 images in the Lake District National Park, of which ca. 850 were rated in the ScenicOrNot project. Such biases are a typical issue in VGI [13], whose handling requires care. Our results were also sensitive to resolution - finer granularities of model reduced model performance (e.g. at 2.5km we could explain 41% of the variation) and coarser granularities increased model performance (e.g. at 10km we could explain 67% of the variation). These results are not unexpected, since firstly the available training data is reduced as resolution becomes finer and, secondly, a coarser model smooths variation and is thus easier to predict, but conveys less fine grained information at the landscape scale.

Our best model used relatively simple features (800 most frequent unigrams, tf-idf and a dictionary of adjectives associated with landscape). Using bigrams, which might be expected to better capture noun phrases associated with scenic locations (e.g. 'pleasant landscape') did not in practice improve model performance, an observation which has been made in other contexts [26]. Verbs of perception appear equally likely to be used in scenic or non-scenic contexts, and were also not useful features in our model.

To our knowledge, our approach is the first attempt to use language to spatially model landscape preference, and it has obvious potential to be combined with other approaches to modelling scenicness based either on user frequentation, physical properties of landscape, or combinations thereof [36, 32, 37].

5 Conclusions and outlook

Our work took advantage of two datasets created by volunteers with very different characteristics. Key to their use in our research were firstly the size and spatial extent of both datasets, and secondly the richness of the textual descriptions associated with Geograph images. Our results demonstrate ways in which VGI and crowdsourcing can allow us to explore questions about how space, and in our case scenicness, is captured through use of language, and demonstrate the potential of such approaches. In particular, we observed:

- clear patterns in the nouns associated with scenicness, suggesting a continuum from heavily developed scenes through more rural to perceived natural scenes. Interpreting and using terms to explain scenicness in isolation is challenging, and we suggest that terms should be analysed in isolation with caution;
- a language-based model can predict some 52% of variance in scenicness, comparable with traditional approaches and state of the art statistical models based on parameters known to correlate with scenicness (e.g. terrain roughness or presence of water). Our approach allows us to capture potentially culturally varying landscape preference through the proxy of language; and
- explained variance was not strongly influenced by the way single users describe landscapes. This makes it unnecessary to restrict the appearance of descriptions of single user either in training or in test datasets.

It is important to note that the approaches we take to modelling scenicness, in contrast to our interpretation of word clouds, essentially use a *bag of words* model, where dependencies between terms are not explicitly modelled. In future research we will explore whether, and how, modelling such dependencies might contribute to our understanding of landscape aesthetics. Importantly, we do not claim that our results are universal, but rather reflect the relationship between landscape and language in a particular cultural setting.

We see this work as an example of the use of textual descriptions to explore culturally determined properties of landscape through language. We also intend to explore the transferability of our results to other user generated content (e.g. Flickr or OpenStreetMap), to other spatial regions and languages (e.g. on mainland Europe) and the impact of including additional spatial data on model performance (e.g. terrain models or land cover data). Furthermore, we see great value in attempting to use the literature to build a taxonomy of scene types, and explore their influence on our model. Such an approach could also take advantage of the "unwritten" parts of our descriptions, for example in terms of the arrangements or presence of objects in a particular image or the relationships between colours through content-based analysis of image content associated with descriptions.

Acknowledgements. We would like to thank Michele Volpi for the fruitful discussions and his and the reviewers helpful comments. We would also like to gratefully acknowledge contributors to ScenicOrNot (Open Database Licence) and Geograph British Isles (Creative Commons Attribution-ShareAlike 2.5 Licence).

References

1 J. Appleton. Prospect and Refuges Revisited. In J. L. Nasar, editor, *Environmental Aesthetics: Theory, Research and Applications*, pages 27–44. Cambridge University Press, 1988.

2 J. A. Benfield, P. A. Bell, L. J. Troup, and N. C. Soderstrom. Aesthetic and affective effects of vocal and traffic noise on natural landscape assessment. *Journal of Environmental Psychology*, 30(1):103–111, 2010. doi:10.1016/j.jenvp.2009.10.002.

3 S. Casalegno, R. Inger, C. DeSilvey, and K. J. Gaston. Spatial Covariance between Aesthetic Value & Other Ecosystem Services. *PLOS ONE*, 8(6), 2013. doi:10.1371/journal.pone.0068437.

4 A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2-3):81–227, 2012.

5 A. Dunkel. Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning*, 142:173–186, 2015. doi:10.1016/j.landurbplan.2015.02.022.

6 A. J. Edwardes and R. S. Purves. A theoretical grounding for semantic descriptions of place. *Proceedings of the 7th international conference on Web and wireless geographical information systems*, pages 106–120, 2007. doi:10.1007/978-3-540-76925-5_8.

7 S. Elwood, M. F. Goodchild, and D. Z. Sui. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, 102(3):571–590, 2012. doi:10.1080/00045608.2011.595657.

8 P. Fisher and D. J. Unwin. Re-presenting Geographical Information Systems. In Fisher, Peter and Unwin, David J., editor, *Re-presenting GIS*, pages 1–14. Wiley Sons London, London, 2005.

9 S. Frank, Ch. Fürst, A. Witt, L. Koschke, and F. Makeschin. Making use of the ecosystem services concept in regional planning-trade-offs from reducing water erosion. *Landscape Ecology*, pages 1–15, 2014. doi:10.1007/s10980-014-9992-3.

10 F. Girardin, J. Blat, F. Calabrese, F. Dal Fiore, and C. Ratti. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7(4):36–44, 2008. doi:10.1109/MPRV.2008.71.

11 G. Gliozzo, N. Pettorelli, and M. Haklay. Using crowdsourced imagery to detect cultural ecosystem services: a case study in South Wales, UK. *Ecology and Society*, 21(3), 2016. doi:10.5751/ES-08436-210306.

12 M. F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007. doi:10.1007/s10708-007-9111-y.

13 M. Haklay. Why is participation inequality important? In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, editors, *European Handbook of Crowdsourced Geographic Information*, pages 35–44. Ubiquity Press, London, 2016.

14 V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305, 2000. doi:10.3115/990820.990864.

15 M. Hunziker, M. Buchecker, and T. Hartig. Space and Place – Two Aspects of the Human-landscape Relationship. *Challenges for Landscape Research*, pages 47–62, 2007. doi:10.1007/978-1-4020-4436-6_5.

16 A. Jenkins, A. Croitoru, A. T. Crooks, and A. Stefanidis. Crowdsourcing a Collective Sense of Place. *Plos One*, 11(4):e0152932, 2016. doi:10.1371/journal.pone.0152932.

17 X. Junge, B. Schüpbach, T. Walter, B. Schmid, and P. Lindemann-Matthies. Aesthetic quality of agricultural landscape elements in different seasonal stages in Switzerland. *Landscape and Urban Planning*, 133:67–77, 2015. doi:10.1016/j.landurbplan.2014.09.010.

18 R. Kaplan and S. Kaplan. *The experience of nature: a psychological perspective*. Cambridge University Press, 1989.

19 A. Lothian. Landscape and the philosophy of aesthetics: Is landscape quality inherent in the landscape or in the eye of the beholder? *Landscape and Urban Planning*, 44(4):177–198, 1999. doi:10.1016/S0169-2046(99)00019-5.

20 A. Lothian. Scenic perceptions of the visual effects of wind farms on South Australian landscapes. *Geographical Research*, 46(2):196–207, 2008. doi:10.1111/j.1745-5871.2008.00510.x.

21 Ch. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999. doi:10.1145/601858.601867.

22 D. M. Mark, A. G. Turk, N. Burenhult, and D. Stea, editors. *Landscape in language. Transdisciplinary perspectives.* John Benjamins, Amsterdam/Philadelphia, 2011.

23 J. Mascaro, G. P. Asner, D. E. Knapp, T. Kennedy-Bowdoin, R. E. Martin, Ch. Anderson, M. Higgins, and K. D. Chadwick. A tale of two "Forests": Random Forest machine learning aids tropical Forest carbon mapping. *PLoS ONE*, 9(1):12–16, 2014. doi:10.1371/journal.pone.0085993.

24 J. L. Nasar. *Environmental Aesthetics: Theory, research and Application*. Cambridge edition, 1992.

25 J. F. Palmer. Using spatial metrics to predict scenic perception in a changing landscape: Dennis, Massachusetts. *Landscape and Urban Planning*, 69(2-3):201–218, 2004. doi:10.1016/j.landurbplan.2003.08.010.

26 B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. *Empirical Methods in Natural Language Processing (EMNLP)*, 10(July):79–86, 2002. doi:10.3115/1118693.1118704.

27 T. Plieninger, S. Dijks, E. Oteros-Rozas, and C. Bieling. Assessing, mapping, and quantifying cultural ecosystem services at community level. *Land Use Policy*, 33:118–129, 2013. doi:10.1016/j.landusepol.2012.12.013.

28 R. S. Purves, A. J. Edwardes, and J. Wood. Describing place through user generated content. *First Monday. Peer-reviewed journal on the internet*, 16(9), 2011. doi:10.5210/fm.v16i9.3710.

29 T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1):1–30, 2009. doi:10.1145/1462148.1462149.

30 D. R. Richards and D. A. Friess. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. *Ecological Indicators*, 53:187–195, 2015. doi:10.1016/j.ecolind.2015.01.034.

31 U. Schirpke, E. Tasser, and U. Tappeiner. Predicting scenic beauty of mountain regions. *Landscape and Urban Planning*, 111(1):1–12, 2013. doi:10.1016/j.landurbplan.2012.11.010.

32 C. I. Seresinhe, H. S. Moat, and T. Preis. Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, page 026581351668730, 2017. doi:10.1177/0265813516687302.

33 C. I. Seresinhe, T. Preis, and H. S. Moat. Quantifying the Impact of Scenic Environments on Health. *Scientific Reports*, 5(Article number 16899):1–9, 2015. doi:10.1038/srep16899.

34 B. Smith and D. M. Mark. Geographical categories: an ontological investigation. *International Journal of Geographical Information Science*, 15(7):591–612, 2001. doi:10.1080/13658810110061199.

35 B. Stadler, R. Purves, and M. Tomko. Exploring the Relationship Between Land Cover and Subjective Evaluation of Scenic Beauty through User Generated Content. In *Proceedings of the 25th International Cartographic Conference*, Paris, 2011. doi:10.5167/uzh-52945.

36 P. Tenerelli, U. Demšar, and S. Luque. Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes. *Ecological Indicators*, 64:237–248, 2016. doi:10.1016/j.ecolind.2015.12.042.

37 B. T. van Zanten, D. B. van Berkel, R. K. Meetemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg. Continental scale quatification of landscape values using social media data. *Proceedings of the National Academy of Sciences*, pages 1–7, 2016. doi:10.1073/pnas.1614158113.

38 B. T. van Zanten, P. H. Verburg, M. J. Koetse, and P. J. H. van Beukering. Preferences for European agrarian landscapes: A meta-analysis of case studies. *Landscape and Urban Planning*, 132:89–101, 2014. doi:10.1016/j.landurbplan.2014.08.012.

39 A. Viberg. The verbs of perception: a typological study. *Linguistics*, 21(1):123–162, 2009. doi:10.1515/ling.1983.21.1.123.

40 B. Yang and T. J. Brown. A Cross-Cultural Comparison of Preferences for Landscape Styles and Landscape Elements. *Environment and Behavior*, 24(4):471–507, 1992. doi:10.1177/0013916592244003.

B

APPENDIX: PUBLICATION 2

Chesnokova, O. and Purves, R.S., 2018. From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text. Applied Geography, 93, 103-111.

PhD candidate's contributions: Developing research idea, annotation, data processing excluding statistical analysis of aggregated distributions of sound experiences in the UK, analysis of the results, creation of maps, writing the draft manuscript and incorporating co-author's feedback.

Applied Geography 93 (2018) 103–111

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text

Olga Chesnokova*, Ross S. Purves

Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

ARTICLE INFO

Keywords:
Sound experiences
User generated content
Natural language processing
Landscape

ABSTRACT

The importance of perception through all the senses has been recognized in previous studies on landscape preference, but data on aural perception, as opposed to the visual, remains rare. We seek to bridge this gap by analyzing texts that describe more than 3.5 million georeferenced images, created by more than 12000 volunteers in the Geograph project. Our analysis commences by extracting and automatically disambiguating descriptions that potentially contain verbs and nouns of sound (e.g. rustle, bellow, echo, noise) and adjectives of sound intensity (e.g. deafening, quiet, vociferous). Using random forests we classify more than 8000 descriptions based on the type of sound emitter into geophony (e.g. rustling wind, bubbling waterfall), biophony (e.g. gulls calling, bellowing stag), anthrophony (e.g. roaring jets, rumbling traffic) and perceived absence of sound (e.g. not a sound can be heard) with a precision of 0.81. Further, we additionally classify these descriptions as negative, neutral and positive using an Opinion Lexicon and GloVe word embeddings. Our results show that sentiment classification gives an additional level of understanding of descriptions classified into different types of sound emitters. We see that geophony, biophony and anthrophony cannot be uniquely classified as positive or negative. Our results demonstrate how text can provide a valuable, complementary to field-based studies, source of spatially-referenced information about aural landscape perception.

1. Introduction and background

What is the contribution of sounds to the way people perceive landscapes? And how can we gather information about such perceptions over large spatial scales? User Generated Content (UGC) has proven to be a suitable source for research questions dealing with such phenomena as people's perception of sense of place (Jenkins, Croitoru, Crooks, & Stefanidis, 2016), conceptualizations of natural features (Derungs & Purves, 2016), olfactory perception (Quercia & Schifanella, 2015), visual perception of landscapes (van Zanten et al. 2016) and assessment of the collective value of protected areas (Levin, Mark, & Brown, 2017). In this study we investigate another subjective phenomenon, namely aural perception of landscapes in UGC, with the underlying future aim of integrating sound information in landscape preference models.

Aural perception is an important constituent in landscape preference assessment (Brown & Brabyn, 2012; Sherrouse, Clement, & Semmens, 2011; Tudor, 2014) and is typically integrated using field surveys (Pilcher, Newman, & Manning, 2009) or laboratory sessions (Benfield, Bell, Troup, & Soderstrom, 2010; Manyoky, Wissen Hayek, Heutschi, Pieren, & Grêt-Regamey, 2014). However, these methods do not allow large regions to be characterized and are time consuming. We assume that aural perception of landscape is present in some written descriptions associated with photographs uploaded by individuals in UGC since photographs have been argued to be a good source of information related to shared experiences of places (Fisher & Unwin, 2005), and sound is one important element of such experiences. The following example vividly illustrates such use of language at an individual level: “If you press your nose to the computer screen, you might just catch the scent of the wild garlic, and if you listen carefully you should hear the song of willow warbler and blackcap.[1]” However, if we wish to analyze such descriptions, then important questions remain with respect to how they can be extracted, how common they are, and what properties they have.

1.1. Sound experiences

Although our sensory experience of nature is by definition multisensory, the visual is often privileged in both research and policy. Thus, despite the introduction of ‘soundscape’, ‘acoustic ecology’ and

* Corresponding author.
E-mail addresses: olga.chesnokova@geo.uzh.ch (O. Chesnokova), ross.purves@geo.uzh.ch (R.S. Purves).

[1] http://www.geograph.org.uk/photo/824881.

https://doi.org/10.1016/j.apgeog.2018.02.014
Received 22 December 2017; Received in revised form 20 February 2018; Accepted 21 February 2018
Available online 05 March 2018

'soundscape ecology' (Southworth, 1969; Schafer, 1993; Pijanowski, Farina, Gage, Dumyahn, and Krause, 2011), aural perception is often of secondary importance in modelling landscape preferences. To relate sound to landscape preference it is important to consider the influence of perceived sound emitters as natural or unnatural (Fisher, 1999), rather than simply decibel values, since we do not hear abstract sounds, but "we hear the way *things sound*" (p. 40 Morton, 2009). Krause (2008), in collaboration with Gage, developed a useful taxonomy for sound emitters in landscape, identifying geophony (non-biological natural sounds), biophony (sounds produced by animals) and anthrophony (human-generated sounds).

Fisher (1999) claims that as soon as we perceive a sound as natural it has a positive aesthetic quality. Thus, similar sounds when perceived as being emitted by a jet engine or a waterfall would be considered unpleasant or "majestically powerful," respectively (p. 28–29 Fisher, 1999). Carles, Barrio, and De Lucio (1999) in their study of sound influence on landscape value note that similar to findings in visual perception, water sounds are typically positively connoted. Furthermore, discordant scenes, for example with positive visual (e.g. a water body) and negative aural cues (e.g. the sound of a busy road) were considered to be especially disturbing. In a series of soundwalks reported on by Pérez-Martínez, Torija, and Ruiz (2018), visitors characterized the sounds of certain emitters as being unpleasant, with, for instance, bird calls dominating, and thus detracting from landscape aesthetics. The negative effects of anthrophony are reported by Pilcher et al. (2009) to be especially important in wild areas, natural parks and other protected areas, where the intrusion of anthropogenic sounds is more disturbing. All of these studies provide us with useful clues as to how aural perception influences landscape perception, but none of them are easily applied across large regions.

1.2. User generated content and extraction of subjective phenomena from language

Our starting point is the hypothesis, based on an initial exploration of content, that UGC can be used to estimate aural perception of landscapes in the British Isles. This hypothesis is supported by previous work which has shown that, for example, tags associated with Flickr images or Tweets content have strong associations with place (Jenkins et al. 2016; Rattenbury, Good, & Naaman, 2007) or that olfactory perception of urban landscapes can be explored through UGC (Quercia & Schifanella, 2015). The same team of researchers also generated maps of urban noises using tags (Aiello, Schifanella, Quercia, & Aletta, 2016) by relating particular terms (e.g. church, car, dog) to particular sounds. However, their study implicitly links sounds to terms without clear evidence of the actual perception of sounds at a location. Similarly, analysis of spectrograms recorded by acoustic sensors (e.g. Pijanowski, Villanueva-Rivera, et al. 2011) does not allow a direct link between the presence of sounds and their perception by humans.

In this paper we build on previous work in two key ways. Firstly, the methods currently used in estimation of aural perception are time consuming and are not suitable for large regions. Using UGC provides an opportunity to explore the link between aural perception and landscapes across the British Isles. Secondly, in the case of recorded sounds presented in laboratory sessions the nature of a sound is abstracted from its context in the landscape. Therefore, we here set out to explore the efficacy of a range of methods for extracting and classifying textual descriptions related to aural perception of sounds, and apply sentiment analysis methods to explore the extent to which landscape descriptions related to different sound emitters can be characterized as positive, neutral or negative. We then explore, quantitatively and qualitatively how aural perception is characterized in our corpus, zooming in to explore local patterns in the description of sound experiences and zooming out to characterize the prominence and distribution of different sound experiences.

2. Data and methods

2.1. Data and study region

As a corpus we used descriptions associated with georeferenced pictures collated through the crowdsourced project Geograph British Isles. Geograph was launched in 2005 with the aim of documenting landscapes through the combination of representative pictures of a location and associated textual descriptions referring to individual grid squares at a granularity of 1 km in Great Britain and Ireland. Geograph contains simple game play elements, with the first contribution to a grid square being awarded more points, and has an active community of more than 12000 users. Similar to most UGC, contributions are biased, with a small number of users[2] contributing the majority of the data, but in previous work it has been shown that descriptions are not strongly biased by individual users, perhaps because of the clear aims and moderation of the uploaded photographs. Furthermore, in a survey carried out by the projects' initiators, users stated that it was important to be sure that the photographs and descriptions are archived for generations to come, and that they be used for educational purposes and promotion of local history. Since no mobile version of Geograph exists we assume that descriptions are written when photographs are uploaded from the desktop computer, though we found evidence that some users take notes in the field.[3] The data used in this paper were downloaded in June 2016, and consisted of more than 5 million photographs, of which more than 3.5 million also had a textual description, and are available under a Creative Commons Attribution-ShareAlike 2.5 License.

2.2. Method overview

Our approach to extracting, classifying and evaluating aural descriptions from the corpus involved three distinct methodological steps:

1. Extraction of descriptions referring to either experienced sounds or perceived absence of sound
2. Classification of the extracted descriptions according to a taxonomy of sound emitters
3. Allocation of sentiment values to each classified description of sound

Fundamental to our work in the first two tasks was the development of an annotated corpus, which was used to evaluate the quality of our extraction rules, and to serve as training and test data for our classifier. Fig. 1 gives an overview of the key steps carried out and described below.

2.2.1. Rules of annotation

As is typical in work on natural language, we created an annotated dataset to, firstly, better understand the properties and use of language in our corpus, secondly, to provide training data for our classifier, and thirdly to evaluate the efficacy of our methods. The annotated dataset contained examples of either descriptions referring to perceived sounds (and thus, not *per se* all detectable sounds) or their perceived absence and we classified these examples according to the type of referenced sound emitter (Table 1).

Descriptions of the following cases were all annotated as related to sound experience:

- aural perception at the moment the photograph was taken, for

[2] Detailed demographic data about users are not available, but based on a survey carried out by the project initiators it appears that users are in general more likely to be over 50 and male.

[3] I made a note on the map that whilst photographing this, the larks were almost deafening! Source: http://www.geograph.org.uk/photo/902702.

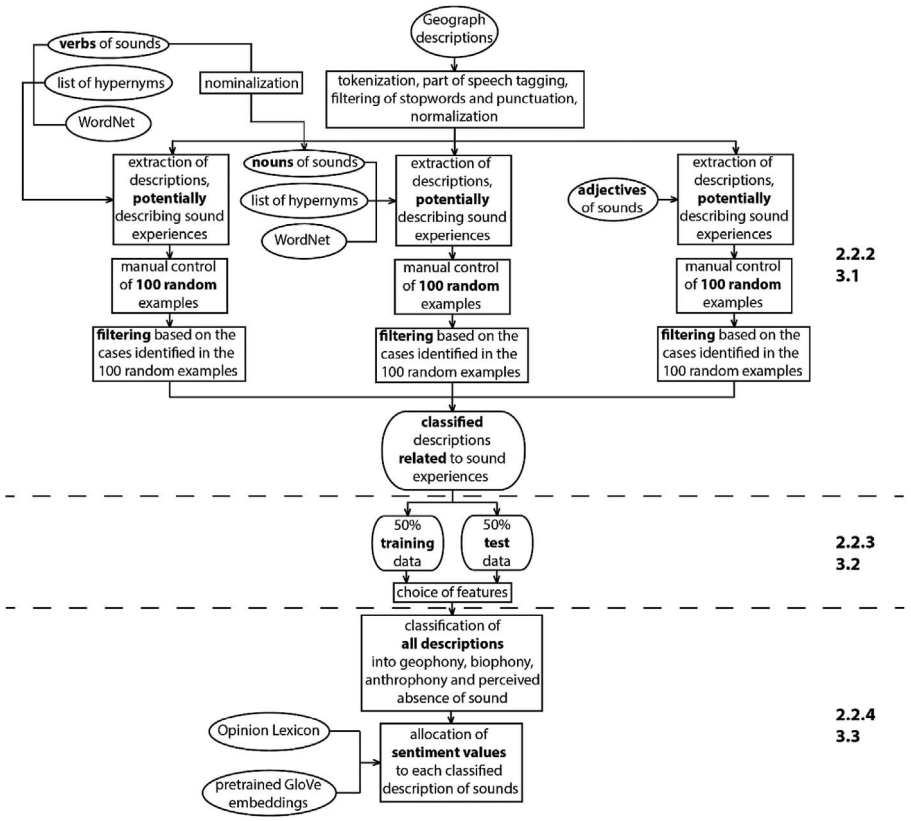


Fig. 1. Steps of data extraction and classification, relevant section numbers are indicated on the right.

Table 1
Types of sound emitters and their description after (Krause, 2008).

Type of sound emitter	Description
Geophony	Descriptions of natural sounds produced by non-biological sources, e.g. wind, waves, thunder, etc. ' … The pieces of ice were building up causing a swishing noise.[a]'
Biophony	Descriptions of sounds produced by animals. ' … If the picture came with sound, you'd hear the constant buzz of insects, the birds singing in the hedges and swifts screaming overhead. … [b]'
Anthrophony	Descriptions of sound produced by humans (including human voices) and anthropogenic objects (e.g. power plant). ' … Aircraft noise is a continual detractor in this intrinsically peaceful countryside.[c]'
Perceived absence of sound	Explicit description of absence of sound, e.g. 'quiet on Sunday morning,' 'not a sound can be heard.' ' … A curious, secret spot, yards away from the thunderous noise of the dual carriageway.[d]'
Mixed	Descriptions including two and more sound emitters, e.g. 'singing birds and roaring traffic' or ' … quiet canal, only the faint hum of the A1 can be heard. … [e]'
Unclear	The sound emitter is unclear, including the references to sound emitter as 'it' or 'they'. 'They look good, but they're noisy![f]'

[a] http://www.geograph.org.uk/photo/233383.
[b] http://www.geograph.org.uk/photo/3507826.
[c] http://www.geograph.org.uk/photo/2035700.
[d] http://www.geograph.org.uk/photo/572030.
[e] http://www.geograph.org.uk/photo/2012725.
[f] http://www.geograph.org.uk/photo/319060.

example, *skylarks are singing*, *running water can be heard*;
- motion of objects described using "sound verbs," e.g. *traffic thunders past*, *the stream gurgles*;
- explicit references to the possibility of sounds (even from the past): *apparently there is a marked echo in the area if one shouts loudly*;
- explicit references to the absence of sound: *the traffic no longer rumbles through their village*;
- aural perception expressed in poems included in the description; and
- indoor sounds.

Descriptions not classified as aural perception included the following:

- with no explicit reference to aural perception: *note the use of straw bales as a noise barrier*;
- of sounds produced by the author of the commentary (e.g. singing, whistling); and
- including similes or metaphors: *the blank walls cry out for some decoration.*

The full annotation was performed by only one person. To test the usefulness of the annotation rules, a second person annotated 100 randomly selected descriptions and inter-annotator agreement was calculated (Landis & Koch, 1977). For annotation of extracted (Cohens Kappa = 0.80) and classified sounds (Cohens Kappa = 0.88) inter-annotator agreement was *almost perfect* according to the classification of

the Landis and Koch (1977), implying that annotation rules used are clear and that the annotation was consistent.

2.2.2. Extraction of descriptions related to sound experiences

We extracted descriptions of sound experiences using a combination of natural language processing methods. To reduce the effects of bias induced by participation inequality, we firstly removed similar descriptions generated by the same user by comparing sequences.

For all remaining descriptions we then carried out part of speech tagging, and using a lexicon of sound verbs extracted candidate sound descriptions after normalizing descriptions by lemmatization. Our initial list of verbs was based on those listed by Levin (1993) as verbs of sound emission, verbs of sounds made by animals and verbs of sound existence. To these verbs we added synonyms extracted from WordNet and clearly related to sound, leading to a total of 196 verbs. Since many of these verbs are polysemous we disambiguated verb usage at sentence level using WordNet hypernyms (categories) associated with the verb and its sentence context using the Lesk algorithm (Manning & Schutze, 1999). We carried out an analogous process for nouns after nominalizing our verb list. Finally, we also extracted descriptions using adjectives contained in a lexicon of sound-related adjectives. However, since WordNet does not contain adjectives in its hierarchy we manually reduced the lexicon of sound-related adjectives[4] to those we judged least likely to be used ambiguously (e.g. we retained *quiet* but not *pleasing*).

Since our rules aim at identifying candidate sound descriptions (i.e. high recall), we implemented them and then annotated a subset of candidate descriptions. Based on the properties of these subsets (i.e. commonly occurring false positives leading to lower precision) we then refined the rules used before annotating the sound descriptions extracted after refinement.

2.2.3. Feature choice

In order to classify descriptions related to sound experiences into types of sound emitter we use random forest classification.[5] Random forests are well suited to classification tasks using diverse feature types, are robust to extraneous features and are straightforward to train (Criminisi, Shotton, & Konukoglu, 2011). Very widely used features in training text classifiers are frequent n-grams – sequences of words found in text (i.e. unigrams are individual words, bigrams are sequences of two words) (Manning & Schutze, 1999). Since our classification task is to identify descriptions related to geophony, biophony, anthrophony and perceived absence of sound we use additional features we judged likely to be useful as described in Table 2. As well as these four classes, we also labelled (and thus trained our classifier on) descriptions belonging to mixed and unclear classes.

2.2.4. Sentiment analysis

The general procedure of allocating sentiment values to a text is described in (Iyyer, Manjunatha, Boyd-Graber, & Daumé , 2015) and was followed here. Firstly, we take an existing general Opinion Lexicon (Hu & Liu, 2004). Though it would be beneficial to have a domain-specific lexicon (Choi & Cardie, 2009), to our knowledge no such lexicon exists in the domain of landscape properties perception. Secondly, using a pretrained set of GloVe word embeddings (Pennington, Socher, & Manning, 2014) we train a gradient descent model[6] to assign polarity values (1 or −1) to all the words we have in our descriptions, and not only those contained in the Opinion Lexicon. Finally, we assign a sentiment value to each description by averaging word sentiment values for a description.

[4] http://www.sightwordsgame.com/parts-of-speech/adjectives/sound/.

[5] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[6] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html.

3. Results and interpretation

3.1. Annotation and extraction of candidate sound descriptions

Annotation was carried out for all descriptions identified as candidate sound descriptions according to the rules described in §2.2.2. Table 3 gives a breakdown of this process, and we summarize important details below.

Based on our initial rulesets, we initially extracted 2436, 5247 and 11453 descriptions based on verbs, nominalized verbs and adjectives respectively. After filtering very similar descriptions and duplicates (i.e. descriptions extracted using our rulesets more than once) a total of 2250, 4730 and 11410 candidate descriptions remained.

For each set we then annotated 100 randomly selected descriptions, and calculated precision. We then used the false positives in each set of candidate descriptions to identify common errors and refined our rules on this basis. For verbs, our initial precision was 0.53. A small number of verbs appeared to be very commonly used polysemously (e.g. *echoes the style of Victorian buildings* or *the house was knocked down*). For a set of five such verbs, we then removed descriptions which contained only these, and no other sound verbs. After this refinement, we extracted 1653 descriptions and annotated all of these. Precision with our refined rules was 0.76.

For nouns, the initial precision was low (0.20) for an annotated random sample of 100 descriptions. A small number of very common polysemous nouns were removed if descriptions contained only these nouns (e.g. *the tree bark is very pretty* or *a clump of bushes is visible on the horizon*), and with the new rules we extracted 1342 descriptions with a precision of 0.68.

Adjectives generated by far the most descriptions, and based on an initial random sample of 100 descriptions precision was 0.56. Since we could not use the Lesk algorithm to disambiguate such adjectives (as hypernyms for adjectives are not contained in WordNet) ambiguity was not considered in our initial extraction. Exploring false positives we noted that *quiet* often appeared to be used in a more general sense to refer to frequency (e.g. *there is not much traffic on this quiet lane*), and added a rule to filter descriptions referring to quiet transport routes using a dependency parser.[7] We thus once again removed descriptions which only contained such phrases and triggered no other rules. Since, in contrast to our verbs and nouns, some 6805 descriptions were extracted, we annotated a random sample of 1000. Based on this sample we achieved a precision of 0.81 for descriptions extracted using adjectives.

After the process of annotation and extraction we created a final corpus of sound descriptions to be used as a training dataset in the classification step. For verbs and nouns we retained only those descriptions which we had annotated as containing sound, while for adjectives these were based on a precision of 0.81. Our complete collection thus contained 8784 descriptions contributed by 1074 unique users. 3036 of the descriptions were annotated.

Table 4 shows the classification of sound emitters according to our annotated corpus. Several points are worthy of note. Firstly, anthrophony is more common than either biophony or geophony. Secondly, mixed and unclear descriptions are relatively rare. Thirdly, geophony and anthrophony appear to be best extracted using a combination of verbs and nouns, while biophony is dominated by the use of verbs. In contrast, absence of sound is characterized by adjectives, reflecting that these descriptions emphasize a property of a location and are not in themselves captured by either verbs or nouns.

3.2. Data classification

Table 5 shows the results of a set of sensitivity tests exploring the

[7] https://spacy.io/usage/linguistic-features#section-dependency-parse.

Table 2
Features used in random forest classifier.

Feature	Description
1: Presence of frequent n-grams	N most frequent uni and bigrams from our corpus after removal of stop words and lemmatization (binary)
2: Presence of British birds and animals	- List of British birds (source: Wikipedia, 198 birds); list of British mammals (source: Wikipedia, 45 mammals) (binary)
3: Presence of transport related terms	Curated list of transport related terms (e.g. train, bus, railway, road, 14 terms) (binary)
4: Presence of natural landscape features and associated qualities	Selected elements based on the list of elements and qualities from Purves, Edwardes, and Wood (2011) (e.g. water, river, sea, hill, fog; 35 terms) (binary)
5: Frequency of references to classified roads	List of all classified roads identified using regular expressions of the form MXX, AXX and BXX where XX are 1 or more digits and M, A and B are motorways, primary and secondary routes (integer)

Table 3
Summary of the steps used to extract descriptions related to sound experiences.

Steps	Based on verbs	Based on nouns	Based on adjectives
Extraction using hypernyms and lists (only lists in the case of adjectives)	2436	5247	11453
After filtering very similar descriptions contributed by the same user	2250	4797	10817
After filtering descriptions already present in the previous dataset	–	4730	11410
Precision of 100 randomly selected examples	0.53	0.20	0.56
After filtering based on the results of the previous step	1653	1342	6805
New precision	0.76	0.68	0.81
Number of descriptions related to sound experiences	1265	909	862 annotated and 5748 unannotated

Table 4
Number of descriptions per type of sound emitter.

Type of sound emitter	Extracted using verbs	Extracted using nouns	Extracted using adjectives	Overall
Geophony	191	134	14	339
Biophony	355	110	60	525
Anthrophony	646	531	107	1284
Absence of sound	25	72	646	743
Mixed	29	41	19	89
Unclear	19	21	16	56
Total annotated	1265	909	862	3036

Table 5
Random forest classifier performance for different sound emitters and feature combinations.

Features (Table 2) Type	1 (200 unigrams)	1 (500 unigrams)	1 (500 unigrams), 2, 3, 4, 5	1 (500 unigrams), 2, 4
Geophony	P = 0.62	P = 0.84	P = 0.84	P = 0.86
	R = 0.34	R = 0.39	R = 0.31	R = 0.37
Biophony	P = 0.56	P = 0.66	P = 0.76	P = 0.74
	R = 0.54	R = 0.59	R = 0.67	R = 0.69
Anthrophony	P = 0.69	P = 0.72	P = 0.73	P = 0.74
	R = 0.81	R = 0.86	R = 0.90	R = 0.89
Absence of sound	P = 0.92	P = 0.91	P = 0.91	P = 0.92
	R = 0.86	R = 0.87	R = 0.86	R = 0.85
Overall	P = 0.70	P = 0.78	P = 0.81	P = 0.81
	R = 0.64	R = 0.68	R = 0.68	R = 0.70

contribution of various features to the classifier's overall performance, and also illustrates performance at the level of individual classes. Our classifier achieved best results (a precision of 0.81) using the 500 most common unigrams, our list of British birds and mammals and our list of natural features and related qualities. Adding transport related terms and named roads did not improve performance. Of note is the relatively high precision achieved for all classes (with values varying between 0.74 and 0.92) and the poor recall for geophony (0.37) implying that some two thirds of such instances were not identified. However, for this task we judge correct classifications (high precision) to be more important than high recall. Further, we concentrate on the four classes of geophony, biophony, anthrophony and absence of sound, because 98% of the descriptions belong to these classes.

Based on these results, we can map spatial distribution of classified sounds both for the whole corpus (Fig. 2) and explore descriptions of perceived sound experiences as extracted from Geograph locally (Figs. 3 and 4). With respect to Fig. 2 a few points are worthy of note. Firstly, the sound experiences extracted correlate with the overall distribution of images (Spearman rank, $r^2 = 0.67$). Secondly, they are dominated by absence of sound (5146) and descriptions of anthrophony (2275). Descriptions related to biophony (832) are less common, and least prevalent are those of geophony (386). These descriptions are also more prevalent in rural areas, and are only weakly correlated with the locations of anthropogenic sounds (Spearman rank, $r^2 = 0.10$ (biophony); $r^2 = 0.09$ (geophony)).

Fig. 3 demonstrates the efficacy of our approach for a rural area in Scotland, encompassing a range of scenic landscapes and a national park, but also traversed by important roads linking urban centers. Biophony is present in a number of descriptions of red deer, as well as the sounds of black grouse calling. Geophony is often related to water, especially *thundering* and *roaring* through gorges and over falls. Anthrophony is most often present in terms of traffic noise, especially where this is heard but not seen. Finally, despite the rural nature of the location, absence of sound, most often in terms of quiet is often reported. Fig. 4 shows results for an area of Central London. Here, geophony is absent completely, and biophony is reported only with respect to naturalized parrots in a park. There are a few references to anthrophony with respect to busy streets and a chiming clock, but the majority of detected references are to absence of sound. As in Fig. 3, these descriptions often contrast the scene with nearby surroundings, or make temporal comparisons (with the photograph taken at a quiet time).

3.3. Sentiment analysis

By calculating sentiment values for classified descriptions we can explore differences in the properties of descriptions, and potentially, the ways in which these are related to perceived environments. Almost 93% of words in our corpus (excluding stop words) were not contained in the Opinion Lexicon, demonstrating the importance of estimating sentiment values using pretrained word embeddings.

To illustrate the use of sentiment analysis in our sound descriptions we stratified sentiment values by generating three relative classes: a negative class consisting of all descriptions with a sentiment value more than half a standard deviation less than the mean, a neutral class of all

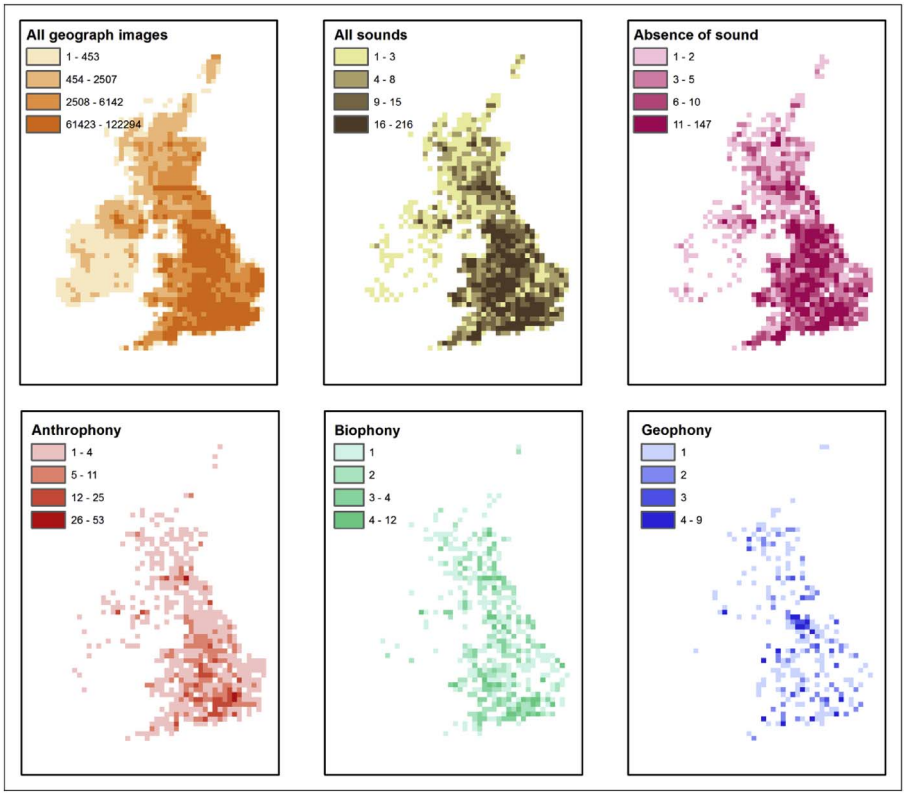


Fig. 2. Aggregated number of descriptions related to sound experiences per type of sound emitter.

Fig. 3. An example of descriptions related to different types of sound emitter of the area between Loch Ness and Cairngorms National Park, Scotland, $n_{desc} = 16$, $n_{users} = 13$. Text source, associated images and authors: www.geograph.org.uk/photo/253824; www.geograph.org.uk/photo/1033986; www.geograph.org.uk/photo/1033836; www.geograph.org.uk/photo/1458692; www.geograph.org.uk/photo/593884; www.geograph.org.uk/photo/432524; www.geograph.org.uk/photo/680910; www.geograph.org.uk/photo/1773441; www.geograph.org.uk/photo/2940613; www.geograph.org.uk/photo/1055504; www.geograph.org.uk/photo/1582508; www.geograph.org.uk/photo/3206512; www.geograph.org.uk/photo/3088559; www.geograph.org.uk/photo/1580898; www.geograph.org.uk/photo/662610; www.geograph.org.uk/photo/1826916.

descriptions with sentiment values lying within half a standard deviation of the mean and a positive class consisting of the remaining descriptions with sentiment values greater than the mean plus half a standard deviation. Fig. 5 shows the distribution of descriptions as a function of their classification. Notable features include the strong association of geophony and biophony with negative descriptions (counter to our naïve expectations) and the association of absence of sound with neutral or positive descriptions. To explore the reasons for these distributions we generated word clouds of the 150 more frequently occurring terms for sentiment values.

Fig. 6 illustrates the resulting word clouds for geophony and biophony respectively. In the negative word clouds for geophony, many weather related words such as *thunder*, *rain*, *wind*, *gale* and *storm* are present. These were not present in the Opinion Lexicon, but have been assigned negative values due to their relationship with other words in the training data, presumably relating negative experiences to weather. *Thunder*, *rain* and *storm* are also prominent in the positive word cloud, along with other terms such as *rainbow*, *waterfall* and *sun*. Associated with negative biophony are many different animals and birds, together with *noise* and some types of sound emission (e.g. *hiss* and *bark*). Positive biophony is related to singing birds and wildlife, and appears, as for geophony, to be related to more natural terms associated with pleasant conditions.

However, we are also interested in how perception of sound varies within particular regions, and in Fig. 7 we explored absence of sound within the boundaries of the UK's 15 national parks. In these word clouds we only retained words which were unique to negative or positive sentiment. Negative sentiment with respect to absence of sound

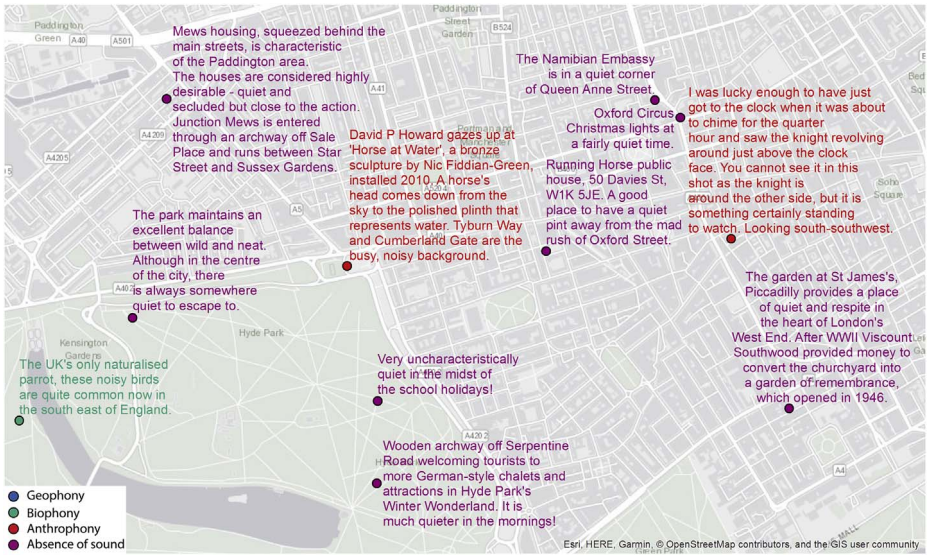


Fig. 4. An example of descriptions related to different types of sound emitter of London, England, $n_{desc} = 11$, $n_{users} = 10$.
Text source, associated images and authors: www.geograph.org.uk/photo/4760309, www.geograph.org.uk/photo/4646159; www.geograph.org.uk/photo/4270494; www.geograph.org.uk/photo/527664; www.geograph.org.uk/photo/2548274; www.geograph.org.uk/photo/1325858; www.geograph.org.uk/photo/1628770; www.geograph.org.uk/photo/119667; www.geograph.org.uk/photo/1999350; www.geograph.org.uk/photo/1661004; www.geograph.org.uk/photo/4418617.

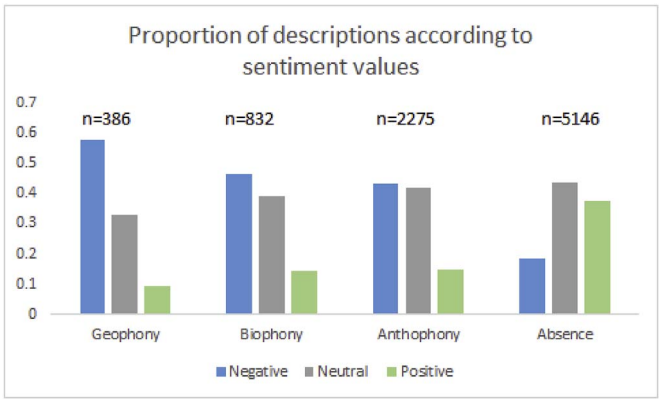


Fig. 5. Proportion of descriptions according to sentiment values.

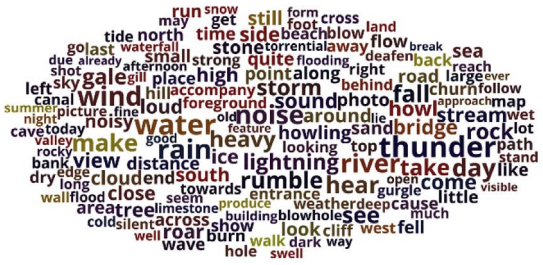
appears to often be related to human activities (e.g. *pump*, *pub*, *railway*, *work*) as well as traffic and isolation (e.g. *traffic*, *backwater*). By contrast, terms relating to positive absence of sound often relate to positively connoted adjectives (e.g. *tranquil*, *enjoy*, *peaceful*, *attractive*, *lovely*) and contain more natural landforms (e.g. *beach*, *summit*, *bay*, *pass*).

4. Discussion

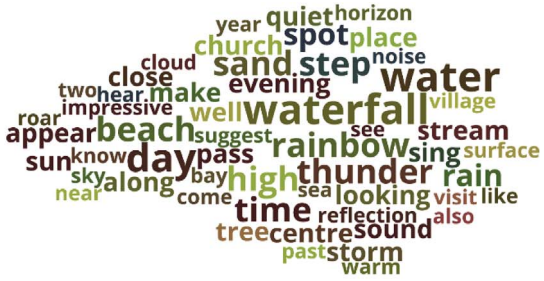
In the following, we discuss our results from two, contrasting, perspectives. Firstly, we explore our methodological contribution, setting out strengths and weaknesses of our approach to the extraction and classification of sound experiences, and the use of sentiment analysis methods on these descriptions. Secondly, we explore our results in the context of previous research on sound experiences both through traditional approaches in landscape research and research based on extraction of perception through UGC.

Our first important contribution is the creation of an annotated, classified corpus of sound descriptions consisting of 8784 descriptions associated with georeferenced images. Creating this corpus would not have been possible without the use of our heuristic methods to extract these descriptions, which were iteratively developed and have a mean precision of 0.75. Our heuristics, based on sound-related lexicons and, in the case of verbs and nouns, disambiguation using hypernyms and the Lesk algorithm are thus sufficiently accurate to allow us to reliably extract sound descriptions from a large collection. However, as is often the case in such work, we have no knowledge of the recall of our method, since this would require us to annotate by hand a very large volume of descriptions. In this particular case, because sound descriptions are rare (making up around 0.25% of our corpus in total) we would have to annotate some 400 descriptions to find a single sound description, and such a manual approach would be prohibitively time consuming. Further to creating our corpus of sound descriptions, we trained a classifier to allocate these to the classes from the taxonomy proposed by Krause (2008). Our best performing set-up used the 500 most frequent unigrams, presence of British birds or mammals and presence of natural features and related qualities in descriptions and achieved a precision of 0.81. However, here we were also able to estimate recall, since our classifier ran on annotated examples of sound experiences. Although overall recall was excellent (0.70) we note that in the case of geophony our classifier performed less well, with a recall of only 0.37. The most likely explanation for this poor performance is the low number of examples of geophony overall, resulting in limited training data for the classifier, especially when compared to anthrophony and absence of sound. However, it is important to note that our approach gives high precision – in other words though not all examples of geophony are classified, those that are, are typically correctly classified. To carry out sentiment analysis we used an Opinion Lexicon to assign values to every non-stop word in a description. Since only around 7% of the words in our descriptions were contained in the lexicon, we used word embeddings and a gradient descent model to assign polarities to the remaining 93% of words. It is important to note that the polarities in the original lexicon are based on general connotations of words with positive or negative polarity, and not those specific to landscapes. Thus, *wild, mystery* and *frozen* all have negative polarities, although all of these terms might be associated with positive values in landscape terms. For example, mystery is suggested as a predictor of environmental preference (Kaplan & Kaplan, 1989). Our approach demonstrates how sentiment analysis can be used to stratify aural descriptions, and as shown in Figs. 6 and 7, to generate interpretable summaries of some landscape properties in terms of sounds and preferences.

Methodologically, our approach has a number of limitations. Firstly, our methods have been developed on a specific collection, and although the rules are general, they have not been tested on other corpora. Nonetheless, by privileging precision over recall, we are reasonably confident that the approach taken should work on other, similar corpora. Secondly, our methods are dependent on annotated data, and annotating is challenging even for humans. Thus, despite good inter-annotator agreement, some cases, especially those describing silence and/or quietness are ambiguous with respect to whether the silence is



Negative geophony (n_{desc}=223)



Positive geophony (n_{desc}=36)

Negative biophony (n_{desc}=386)

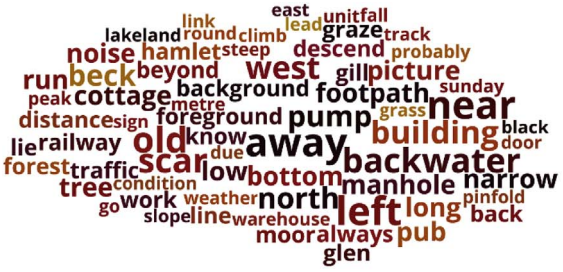
Positive biophony (n_{desc}=119)

Fig. 6. 150 most frequent tokens occurring more than three times for negative and positive descriptions of geophony and biophony.

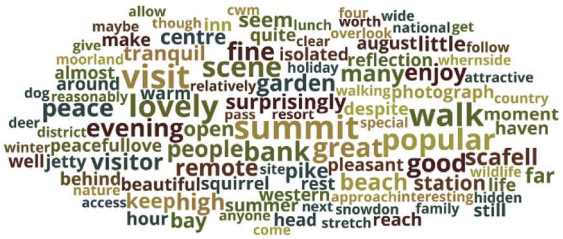
metaphorical or really related to sound. Thirdly, our approach to sentiment analysis is fairly simplistic, and in particular does not currently deal with negation or normalize according to the lengths of descriptions. And finally, we do not take into account dynamic nature of sounds varying over the course of a day or due to changing weather events.

Some of the strengths of our approach are well illustrated by exploring the content in more detail, and comparing it to previous research in this area. Since we generate a large corpus of classified descriptions this is possible, and as illustrated in Figs. 2–4 patterns can be explored both globally and locally. Our results appear to confirm that Krause's (2008) taxonomy is broadly useful, since we find abundant examples of anthrophony, biophony and geophony. However, it is important to note that in our work we take a human-centered approach as in Truax (1978) and Schafer (1993), where we do not concentrate on the whole variety of sounds in the environment but only on sounds perceived and recorded in written descriptions. Furthermore, exploring these patterns at local levels (e.g. Figs. 3 and 4) gives insights into how people perceive landscapes as a combination of all senses. Our methods thus provide a further potential approach in including sound in methods for landscape characterization (c.f. Sherrouse et al. 2011, Brown & Brabyn, 2012).

Sentiment analysis allowed us to stratify descriptions in meaningful ways. Thus, for geophony we observed that adverse weather seemed often to be related to negative descriptions, while positive descriptions, though relatively rare, seemed to be associated with landforms such as waterfalls and beaches, as well as positively connoted weather phenomena such as rainbows. In accord with previous work (e.g. Morton, 2009), our results appear to demonstrate the importance of perceived sound emitters, and we note that for biophony sounds associated with farm (and thus perceived less natural) animals appear to be less well received than wild animals. Further, we also observe that there are differences with respect to sounds related to wild animals, with for example gulls being perceived negatively while skylarks are positive. Similarly, we note that sounds which could be classified as *technophony* (Mullet, Gage, Morton, & Huettmann, 2016) in the anthrophony class appear to be mostly associated with negative sentiment (e.g. words such as *traffic*, *road*, *motorway*, *turbine* are common), while more positively



Negative absence of sound (n_{desc}=121)



Positive absence of sound (n_{desc}=194)

Fig. 7. Most frequent unique tokens occurring more than three times for negative and positive descriptions of absence of sound in the UK's 15 national parks.

associated anthrophony is reflected by words such as *clock*, *bell*, *music* and *sing*, combining *mechanistic* (e.g. chiming clock) and *oral* (e.g. carols singing) sub-classes of anthrophony (Qi, Gage, Joo, Napoletano, & Biswas, 2008). Again, these results are strikingly congruent with previous work (Pérez-Martínez et al. 2018), suggesting that our approach can usefully complement existing approaches to characterizing aural experiences.

5. Conclusions and outlook

Our aim was to explore the potential of textual descriptions associated with georeferenced photographs as a source of information on perceived sounds in landscapes. Since the dataset used was created by more than 12000 contributors, the resulting extracted descriptions provide us with a bottom-up view of the ways in which sounds are described, and give insights into how landscapes are perceived through multiple senses. Although the overall number is a small proportion of the corpus (some 0.25%), in absolute terms we have extracted more than 8000 sound-related descriptions, classified these according to emitters, and explored how the use of descriptions (and thus the perception of landscape in terms of sound) varies at different scales. Furthermore, by applying sentiment analysis we stratified descriptions and explored preferences within different classes of emitter, moving away from, for example, naïve expectations that natural sounds are *per se* positively evaluated.

Methodologically our contribution can be seen in two ways. Firstly, we have created an annotated corpus of classified descriptions which can serve as a basis for further research. Secondly, we have demonstrated how a combination of methods from natural language processing, going beyond simple extraction based on keywords, and taking account of typical linguistic phenomenal such as syntax and polysemy, allow us to extract and classify sound descriptions with high precision. Our approach to sentiment analysis used word embeddings to learn sentiment values for words not contained in our lexicon. Here we note that results are dependent on the lexicon used, and we propose to develop a domain-specific opinion lexicon focussed on landscape.

Our methods have general potential for future work in a number of ways. For example, they can be used to explore change in perceived sounds over time and thus contribute to the digital humanities. Furthermore, by exploring the relationship between aural descriptions and spatially contiguous models of abstract landscape qualities such as wilderness or tranquility the influence of perceived sounds on such properties can be accorded greater importance than is currently the case. Finally, we see great potential for integrating our results into a more general model of landscape preference based on textual analysis.

Acknowledgements

We are very grateful to Barry Hunter and Robin Stott for sharing anonymous results of the Geograph British Isles users' survey, and we would like to gratefully acknowledge all the contributors to Geograph British Isles (Creative Commons Attribution-ShareAlike 2.5 License). OC is also grateful to Lone Kristensen and Veerle Van Eetvelde for feedback on early versions of this work. Finally, we thank the reviewers for their useful suggestions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apgeog.2018.02.014.

References

Aiello, L. M., Schifanella, R., Quercia, D., & Aletta, F. (2016). Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science, 3*, 150690.

Benfield, J. A., Bell, P. A., Troup, L. J., & Soderstrom, N. C. (2010). Aesthetic and affective effects of vocal and traffic noise on natural landscape assessment. *Journal of Environmental Psychology, 30*(1), 103–111.

Brown, G., & Brabyn, L. (2012). The extrapolation of social landscape values to a national level in New Zealand using landscape character classification. *Applied Geography, 35*(1–2), 84–94.

Carles, J. L., Barrio, I. L., & De Lucio, J. V. (1999). Sound influence on landscape values. *Landscape and Urban Planning, 43*(4), 191–200.

Choi, Y., & Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. *EMNLP '09 proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 590–598). .

Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision, 7*(2–3), 81–227.

Derungs, C., & Purves, R. S. (2016). Characterising landscape variation through spatial folksonomies. *Applied Geography, 75*, 60–70.

Fisher, J. A. (1999). The value of natural sounds. *Journal of Aesthetic Education, 33*(3), 26–42.

Fisher, P., & Unwin, D. J. (2005). In D. J. Fisher, & Unwin (Eds.). *Re-presenting geographical information systems* (pp. 1–14). London: Wiley Sons London.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04: Vol. 4*, (pp. 168–). .

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé, H., III (2015). Deep unordered composition rivals syntactic methods for text classification. *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*.

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PLoS One, 11*(4), e0152932.

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective.* Cambridge University Press.

Krause, B. (2008). Anatomy of the soundscape. *Journal of the Audio Engineering Society, 56*(1/2).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Levin, B. (1993). *English verb classes and alternations.* University of Chicago Press.

Levin, N., Mark, A., & Brown, G. (2017). An evaluation of crowdsourced information for assessing the importance of protected areas. *Applied Geography, 79*, 115–126.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing.* The MIT Press.

Manyoky, M., Wissen Hayek, U., Heutschi, K., Pieren, R., & Grêt-Regamey, A. (2014). Developing a GIS-based visual-acoustic 3D simulation for wind farm assessment. *ISPRS International Journal of Geo-Information, 3*(1), 29–48.

Morton, T. (2009). *Ecology without nature: Rethinking environmental aesthetics.* Harvard: Harvard University Press.

Mullet, T. C., Gage, S. H., Morton, J. M., & Huettmann, F. (2016). Temporal and spatial variation of a winter soundscape in south-central Alaska. *Landscape Ecology, 31*(5), 1117–1137.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Conference on empirical methods in natural language processing (EMNLP 2014)*.

Pérez-Martínez, G., Torija, A. J., & Ruiz, D. P. (2018). Soundscape assessment of a monumental place: A methodology based on the perception of dominant sounds. *Landscape and Urban Planning, 169*, 12–21.

Pijanowski, B. C., Farina, A., Gage, S. H., Dumyahn, S. L., & Krause, B. L. (2011a). What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape Ecology, 26*(9), 1213–1232.

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., et al. (2011b). Soundscape Ecology: The science of sound in the landscape. *BioScience, 61*(3), 203–216.

Pilcher, E. J., Newman, P., & Manning, R. E. (2009). Understanding and managing experiential aspects of soundscapes at muir woods national monument. *Environmental Management, 43*(3), 425–435.

Purves, R. S., Edwardes, A. J., & Wood, J. (2011). Describing place through user generated content. *First Monday, 16*(9) Peer-Reviewed Journal on the Internet.

Qi, J., Gage, S. H., Joo, W., Napoletano, B., & Biswas, S. (2008). Soundscape characteristics of an environment: A new ecological indicator of ecosystems health. In W. Ji (Ed.). *Wetland and water resource modeling and assessment* (pp. 201–211). New York: CRC Press.

Quercia, D., & Schifanella, R. (2015). Smelly Maps: The digital life of urban smellscapes. *9th international AAAI conference on web and social media*.

Rattenbury, T., Good, N., & Naaman, M. (2007). Towards automatic extraction of event and place semantics from Flickr tags. *Proc.of SIGIR: 07*, (pp. 103–110).

Schafer, R. M. (1993). *The soundscape: Our sonic environment and the tuning of the world.* Inner Traditions/Bear & Co.

Sherrouse, B. C., Clement, J. M., & Semmens, D. J. (2011). A GIS application for assessing, mapping, and quantifying the social values of ecosystem services. *Applied Geography, 31*(2), 748–760.

Southworth, M. (1969). *The sonic environment of cities. Environment and behavior* (June) 49–70.

Truax, B. (1978). *Handbook for acoustic ecology.* Burnaby, B.C. Canada: ARC Publications.

Tudor, C. (2014). *An approach to landscape character assessment* (October)56.

van Zanten, B. T., van Berkel, D. B., Meetemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016). Continental scale quantification of landscape values using social media data. *Proceedings of the National Academy of Sciences*, 1–7.

C

APPENDIX: PUBLICATION 3

Chesnokova, O., Taylor, J.E., Gregory, I.N., and Purves, R.S., 2019. Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District. International Journal of Geographical Information Science, 33:12, 2430-2454.

PhD candidate's contributions: Co-developing research idea, annotation, data processing and analysis excluding micro-analysis and statistical tests related to spatial distribution of tranquillity, creation of maps, writing parts of the draft manuscript and incorporating co-authors' feedback.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2019, VOL. 33, NO. 12, 2430–2454
https://doi.org/10.1080/13658816.2018.1552789

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District

Olga Chesnokova [a], Joanna E. Taylor [b], Ian N. Gregory[c] and Ross S. Purves[a]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Presidential Academic Fellow in Digital Humanities, University of Manchester, Manchester, UK; [c]Department of History, Lancaster University, Lancaster, UK

ABSTRACT

We analyse silence and tranquillity in historical and contemporary corpora to understand ways landscapes were—and are—perceived in the Lake District National Park in England. Through macro and microreading we develop a taxonomy of aural experiences, and explore how changes to categories of silence from our taxonomy—for instance, the overall decline in mentions of absolute silence—provide clues to changes in the landscape and soundscape of the Lake District. Modern authors often contrast silence with anthropogenic sounds, while historical authors adhere to a cultural construction where the Lake District is presented as a tranquil area by ignoring industrial sounds. Using sentiment analysis we show that silence and tranquil sounds in our corpora are, as a whole, more positively associated than random text from the corpora, with this difference being especially marked in contemporary descriptions. Focusing closely on individual texts allows us to illustrate how this increased positivity can be related to the emergence of silence and tranquillity as valuable components of landscape. Mapping our corpora confirmed the influence of Wordsworth's writing on descriptions of silence; and revealed the co-location of pockets of tranquillity near to transport arteries in contemporary descriptions.

ARTICLE HISTORY
Received 4 May 2018
Accepted 22 November 2018

KEYWORDS
Landscape; tranquillity; temporal changes; text analysis

1. Introduction

Human perception is shaped by our senses: the ways in which we experience the world are driven by what we see, hear, smell, touch and taste. The importance of this multi-sensory perception of landscape is emphasised in policy documents; the European Landscape Convention, for instance, defines landscapes as 'an area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors' (Council of Europe 2000, p. 2). A diverse range of academic fields—including human geography, landscape ecology, history and computer science—have increasingly recognised the need to move away from considering only what we see, to other ways of perceiving landscapes (e.g. Smith 1994, Pijanowski *et al*. 2011, Quercia and Schifanella 2015). This paper joins this

CONTACT Olga Chesnokova olga.chesnokova@geo.uzh.ch

new tradition of scholarship. From the perspective of spatial computing for the digital humanities, we will demonstrate how blending research methods from literary studies, corpus linguistics and geographical information science—with inputs from disciplines including geography and history—can offer fresh perspectives on landscape change with consequences for the ways we understand a location's development and its social, cultural and ecological status. As we will show, studying landscape(s) is an inherently interdisciplinary endeavour.

In purely spatial terms, landscapes matter because they are heterogeneous: different characteristics make different landscapes special, and drive the desire to protect unique or exemplary landscapes (Tudor 2014). Practically speaking, landscape characterisation based on perception typically combines existing spatial data and implicitly links this to—usually visual—perception. The senses which are thus modelled are always, by definition, incomplete. First, they reflect a particular set of cultures and practices where specific elements are considered more important in, for instance, planning decisions. Second, our ability to conceptualise, abstract and model spatially variable processes in technologies like GIS is limited. Some perceived elements of landscapes are, at least over short intervals of time, easier to conceptualise and model by knowing something about the constituent physical make-up of a scene: at their simplest, perceivable boundaries in a landscape indicate land use or land cover (Turner 2006); viewsheds (e.g. Lake *et al*. 1998, Fisher *et al*. 2004) provide a route to landscape vistas; and coherence can be modelled through shape and distribution of co-occurring parcels (e.g. McGarigal and Marks 1994). Such approaches provide a starting point, at least, for generating plausible spatial models that reflect what might be sensed at some locations. However, landscapes are not static; they are subject to both natural and anthropogenic processes which in turn lead to change.

Understanding and documenting landscape change is an important task in landscape planning, since it provides a baseline with respect to both objective and subjective notions of landscape over time. Despite a recognition that GIS can capture change—either through snapshots or process-based models (Grenon and Smith 2004)—in reality models of landscape change used for practical purposes focus almost exclusively on snapshots at particular moments in time. Even where we take an apparently simple approach to measuring change in the recent past, multiple challenges arise: changes in sensors and their capabilities (Sexton *et al*. 2016); changes in ontologies (Comber *et al*. 2005); cultural and linguistic differences (Burenhult and Levinson 2008); and changes in computational methods and representations (Vasconcelos *et al*. 2002) all affect landscapes over time. As we go further back in time to map historical landscapes, these challenges become more pronounced; primary data describing physical landscapes become increasingly scarce and the complexity of relating historical representations to current data models increases.

Nonetheless, such approaches typically assume that some basic spatial data exist, such as in the form of cartographic products (Leyk *et al*. 2006, Fuchs *et al*. 2015), aerial photographs or satellite imagery (Van Den Berghe *et al*. 2018). Historical Landscape Characterisation—a technique pioneered in England that seeks to model and interpret landscapes with respect to their historical development (Turner 2006, Fairclough and Herring 2016)—is an example of such an approach. Here, the importance of perception in understanding and representing past landscapes in mappable forms is clear. Yet, this process of interpretation is carried out by experts, and has thus been criticised in the context of Landscape Character Assessment as being dominated by values attributed

through 'objective' outsiders (Butler 2016). This criticism, and the relating tension between what are viewed as positivist or oversimplified mappings of landscapes on the one hand, and a lack of pragmatism and contribution to real societal needs to describe and monitor change on the other, has multiple parallels with well-known debates in Geographic Information Science (Pickles 1995, Rundstrom 1995).

We do not propose to rehash these debates here, but rather to try to find some middle ground. As a starting point, we understand written accounts as a window to perception. We use a blend of interdisciplinary methods to analyse text and thus show how landscape is—and was—perceived. We focus on one particular landscape: the English Lake District, a UNESCO World Heritage Site in the North West of England, and one of the most comprehensively recorded landscapes in the world (Nomination 2015). Indeed, the UNESCO designation recognises the region as a 'cultural landscape', so significant has been the effect of art and literature on its historical and contemporary character.

More specifically, what we are interested in here is the perception of a particular rural soundscape. Written sources are particularly important for providing a route to understanding perceptions of experiences which are, otherwise, ephemeral. We explore what written accounts of personal experiences might reveal about the role of sound and, in particular, silence in response to (perceived) natural landscapes. We will see that texts preserve, however imperfectly, sounds which risk being forgotten as social, cultural and technological contexts change (Lowenthal 1976, Smith 1994, p. 233). Although written descriptions do not necessarily reflect the objective soundscape, they do offer insights into personal experiences of place. More than that, they indicate something of the changing social and cultural status of sounds. As we demonstrate, written accounts can provide us with one—admittedly incomplete, yet nevertheless significant—way of understanding what people describe when communicating about both contemporary and historical landscapes, and why this matters for the development of these locations. If these descriptions are about the same places, at different times, then we can pose an important question: can we use written accounts to characterise changes in perceived sounds and silences in landscapes across both space and time?

To answer this question, we first explore how sound and soundscapes have emerged as important components of landscape studies. We then discuss the emergence of notions of tranquillity and quiet at the turn of the eighteenth century, and show how these ideas continue to influence modern-day values associated with rural peace, silence and tranquillity. We introduce our study area, the Lake District, and the two corpora on which our study is based. Finally, we explain the interdisciplinary approach to text analysis which has allowed us to offer conclusions about the nature of tranquillity, and to understand changes to the ways in which the Lake District landscape is valued.

2. Background

2.1. *Sound and silence*

Sound affects us more consistently, perhaps, than any other sense; as Bruce R. Smith observes, we are 'surrounded – and filled – by a continuous field of sound' (1999, p. 9). Yet, it is only comparatively recently that scholarship has begun to reflect on the importance of multisensory perception to human understandings of place and space.

The new awareness of sound represented by the work of scholars like Susan J. Smith, Mark Smith and Alain Corbin (Corbin 1986, Smith 1994, 2004) is indebted to R. Murray Schafer's work on the World Soundscape Project in the 1970s and 1980s. Schafer was predominantly interested in the components which made up an area's soundscape, but more recent work has moved on from the 'acoustic ecology' promoted by the World Soundscape Project (Truax 1978), preferring instead to highlight a 'soundscape ecology' that focuses on interactions between different acoustic elements and their environment (Pijanowski *et al*. 2011). This work delivers new assessments of the ways that humans interact with—or are affected by—the soundscapes they encounter. Specifically, as Pijanowski and his co-authors suggest, more research is needed into 'how natural sounds influence the development of individuals' sense of place, place attachment, or connection to nature', as well as the factors which 'affect human (in)tolerance of soundscape changes, especially where those changes increase noise' (2011, p. 209).

As these scholars imply, acoustic experiences are as subject to contemporary fashion as visual ones. Peter Coates puts it neatly when he writes that '[j]ust as beauty is in the eye of the beholder, noise frequently resides in the ear of the listener' (Coates 2005, p. 641). In short, what is interpreted as a sound—rather than as noise—changes over time. As Isabelle Bour suggests, such transitions can occur slowly over centuries, or as quickly as the course of a day: 'a buskers trumpet,' she concludes, 'is likely to be perceived as sound at midday but as noise at midnight' (Bour 2016). The way a sound is interpreted depends on the characteristics of the sound, including its volume, as well as the affective response it initiates in the listener (MacFarlane *et al*. 2004, p. 134).

The late eighteenth century catalysed this emergence of listening as what Sophia Rosenfeld calls 'a cultural effect as much as a physiological one' (Rosenfeld 2011, p. 318). Debates about what constituted sound, and what noise, emerged at this moment—from when the earliest texts on which we focus here date—when philosophers began developing an acoustic aesthetics that paralleled the development of the picturesque for visual phenomena (Agnew 2012, Joy 2014, Donaldson *et al*. 2017). Thinkers such as William Duff and James Beattie agreed that the combination of certain sounds formed inherent harmonies, and so might produce pleasurable emotional responses in the listener (Dubois 2016). Others, though, grated on the listener; to describe an acoustic experience as 'noisy' has always been pejorative. Paul Hegarty explains that noise is negative because it 'can never be positively, definitely and timelessly located' because it is emblematic of elements that society wishes to resist (Hegarty 2007, p. ix).

This emergence of sound and noise as important topics for discussion is perhaps unsurprising in an age that witnessed profound changes to its soundscapes. Noise control was a prominent concern, in urban environments especially, throughout the eighteenth century. By the Victorian period, city soundscapes had intruded into the countryside: the railway screeched and rattled over the new lines that criss-crossed the country; cacophonies from emerging industrial centres echoed around the surrounding area; and, in towns and cities, new forms of making sound and noise contributed to a sonic shift with profound cultural consequences (Picker 2003, p. 5). Even before the motor car's arrival, the nuances of the soundscape—and, particularly, the delicacies of natural sounds—had been largely masked by human noises. The result was a shrinking of what Bruce Smith calls the 'acoustic horizon'; modern sounds, especially the low drone of the internal combustion engine, obscure other low frequency sounds and

dramatically reduce the distance at which all sounds can be heard by the human ear (1999, p. 51). Smith believes that quiet environments, away from the drone of traffic and hubbub of daily life, expand our acoustic horizons and enable deeper, more meaningful connections between the self and the world (Smith 1999, p. 74).

By the turn of the eighteenth century, urban residents in particular began to long for respite from the constant din of the modern world, and to seek out locations that still maintained—at least relatively—nuanced, natural soundscapes. Following the rapid expansion of industrialisation towards the end of the eighteenth and into the beginning of the nineteenth centuries, it did not take long for cultural evaluations of quietness, as well as noise and sound, to shift. Before then, quietness had seemed to indicate a lack of civilisation; it was a marker of what John Fisher calls 'untrammeled nature', and the impulse was to tame it (1999, p. 27). But as towns and cities grew noisier, those who had the financial means began to seek silence. By the mid-nineteenth century, 'quiet' became equated with 'peace' (Ammer 2013), and the remoteness offered by less accessible regions—such as the Lake District—began to be valued as much for the respite they could offer from the urban din as for their picturesque beauty.

As we will see, contemporary landscape preferences maintain this need for peace and quiet; tranquillity requires sounds and silences that generate states of repose. The search for peace and quiet away from urban centres is not merely a desire for a break from day-to-day pressures; it is also a necessary part of the human requirement for connection with the natural world, and for related feelings of calm. Indeed, one of the main reasons people today give for visiting rural landscapes is the search for peace and quiet; in a 2001 survey conducted by the UK's Department for Environment, Food and Rural Affairs, 'tranquillity' was cited by 58% of respondents as their main motivation for spending time in the countryside, ahead of scenery (46%), open space (40%), fresh air (40%), or plants and wildlife (36%) (MacFarlane *et al*. 2004, p. 7). National organisations charged with the care of these landscapes have increasingly recognised the importance of sound, as well as sight, on visitor experiences (Miller 2008): in the case of the American National Park Service (NPS), for instance, the preservation of soundscapes was enshrined in an act signed by President Woodrow Wilson in 1916. Today, the NPS runs a comprehensive programme to promote the conservation of soundscapes in their care, and to educate their visitors about the importance of natural sounds as 'part of a web of resources [that are] vital to park ecosystems'.[1]

In Britain, the protection of the Lake District's tranquillity—and the 'sense of space and freedom' it engenders—was one of the motivating factors behind the Lake District National Park's successful application for UNESCO World Heritage Site status in 2017 (Nomination 2015, 2.72). This designation, the historical association of the region with peace and quiet, and the volume of writing it has inspired make it an exemplary location to test the potential of our approach.

3. Writing Lakeland: the corpus of Lake District writing and Geograph

The boundaries of the Lake District National Park (Figure 1) were established in the mid twentieth century, but the region has been celebrated for its picturesque beauty, opportunities for outdoor activities (including mountain climbing and wild swimming) and comparative remoteness since the late eighteenth century. It is particularly

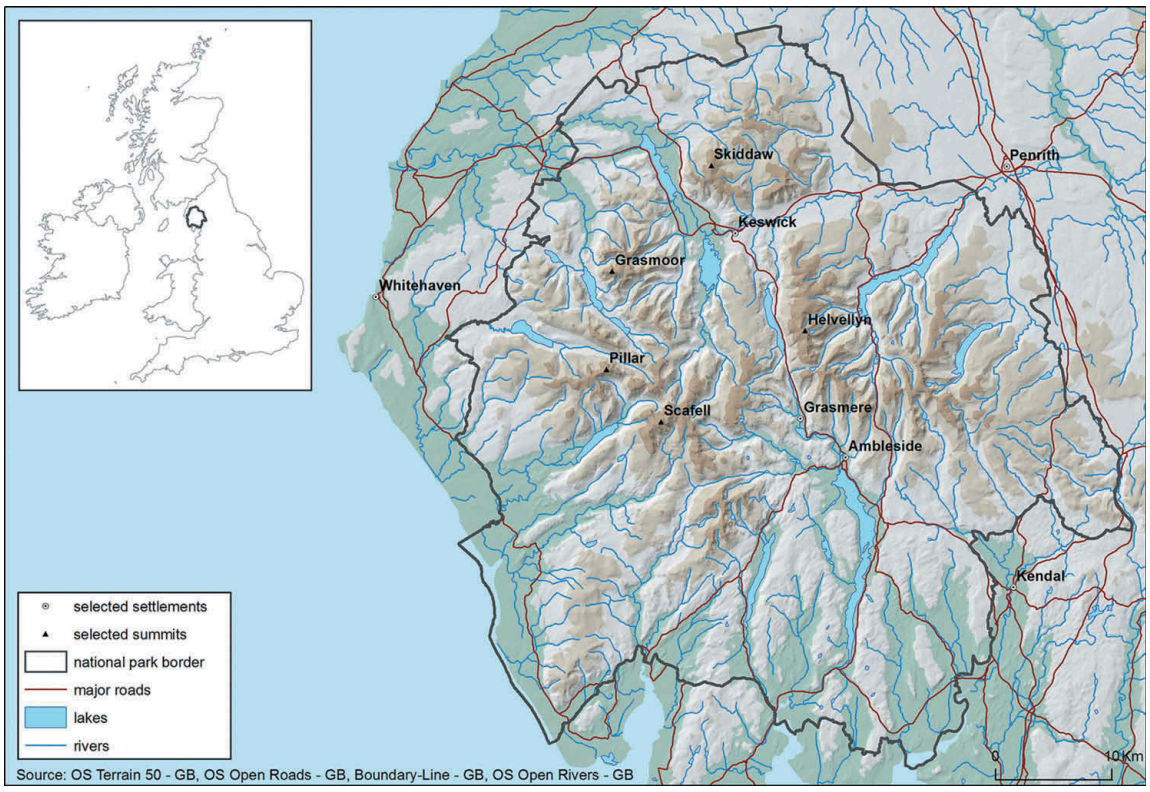


Figure 1. Our study area: The Lake District in North West England.

renowned for the authors and artists who have made their homes there, or found inspiration in its dramatic upland landscapes: William and Dorothy Wordsworth, John Ruskin, Beatrix Potter and Arthur Ransome all lived in the area, and their writing and wider agricultural, geological and artistic interests have—in different ways—shaped today's Lake District. Alongside these famous figures, a host of lesser-known writers and the increasingly multitudinous voices of visitors to the region have influenced the conservation practices that maintain this area as a particular kind of cultural landscape (Nomination 2015, Donaldson *et al*. 2017).

We are especially interested in written responses to the Lake District from the eighteenth century to the present day extracted from two sources: the historical Corpus of Lake District Writing (CLDW)[2] and the contemporary Geograph Project.[3] The Corpus of Lake District Writing is a georeferenced collection of writing about the Lake District and the surrounding area; broadly speaking, it is interested in the modern county of Cumbria, which was formed in the 1970s by merging the old counties of Cumberland and Westmorland with the Lancashire Hundred of Lonsdale North of the Sands. Currently, it comprises 1.5 million words, consisting of 80 texts that include novels, poetry, epistolary fiction, non-fictional essays, topographical accounts and travel writing about the region (Donaldson *et al*. 2015, Gregory and Donaldson 2016, Murrieta-Flores *et al*. 2017). To focus as closely as possible on lived experiences of the historical Lake District's soundscape, for this project we selected only the non-fictional prose accounts as the most similar to our second corpus. The version of the CLDW we discuss in this article, then, contains 61 texts written by 55 authors, and comprises 1.3 million words.

Our second corpus is constituted of text from the crowdsourced project Geograph British Isles. Geograph, launched in 2005, aims to document landscape at the scale of

1km grid squares. The site combines photographs with textual descriptions, written by the photographers themselves, and has an active community of more than 12,000 users. Like most User Generated Content platforms, many of Geograph's contributions tend to be made by a small number of users, most of whom are men over the age of 50.[4] For the Lake District and its immediate surroundings, there are 90,705 photographs, contributed by 1218 authors. Of these, 64,795 include descriptions of photographs, which total 1.4 million words written by 1076 authors. Geograph is well suited to our task since, firstly—and unlike other photo-sharing platforms like Flickr—images are associated with full text descriptions rather than tags which are better suited to semantic analysis. Secondly, geographic coverage is not strongly biased to urban areas because of the use of a points system that motivates documentation of new grid squares (Antoniou *et al*. 2010). Thirdly, Geograph explicitly focuses on collecting descriptions of experienced (and thus perceived) landscapes (Chesnokova and Purves 2018). Example extracts from both the CLDW and Geograph can be found in Figures 5 and 6.

Both corpora are composed by authors who, by the nature of the genres in which they write, are highly aware of the relationship between location and description. Comparing these sources therefore allowed us to assess a diachronic sweep of writing about the region, and to trace in it evolving understandings of the Lake District's status as an enclave for tranquillity in an increasingly industrialised, mechanised and digitised world. Nonetheless, there are differences between the two corpora. Firstly, and most importantly, the texts in each were written with very different aims and audiences in mind. Nevertheless, both corpora contain references to first-person perceptions of silence, and document something of the lived experience of listening to the landscape. Secondly, Geograph has significantly more individual authors, and thus has the potential for a wider range of perspectives (although, as we will see shortly, their approaches are remarkably homogeneous). Thirdly, the sentence length in the CLDW is—unsurprisingly, given the dates of composition of the texts—almost twice as long as in the modern corpus, and these are much more heavily punctuated. Nevertheless, with 43,269 sentences (for the CLDW) and 98,206 (for Geograph) sentences each, our corpora are of a comparable size.

4. Finding the middle ground

4.1. *Overview of the process*

To identify, compare and analyse text relating to sounds in these two corpora, we combined so-called macro and microanalysis of the texts (Jockers 2013). We iteratively applied a mixture of text analysis methods, guided and informed by the contemporary acoustic-cultural contexts in which the texts were written, to process the works quantitatively and to identify features for further exploration. By also qualitatively microanalysing the texts in our corpora, we were able to ask more specific questions about the relationship between personal perceptions of landscape and soundscape, and to develop a more detailed understanding of how the soundscape affects individuals' senses of peace and tranquillity in the Lake District National Park. Finally, we mapped a subset of texts from Geograph and the CLDW to explore spatial patterns in the extracted texts.

Our approach can be grouped into four main tasks, listed here in the order undertaken in our methodological pipeline (thus, e.g., annotation was carried out after an initial exploration of the corpora):

- Preprocessing: a set of basic methods to prepare a text snippet for analysis.
- Search and comparison: extraction of texts potentially describing sounds and comparison of their associated content within and between corpora.
- Annotation: classifications of the nature of silence and its emitters in text snippets.
- Enhancement: the calculation of additional properties related to text snippets, for example in our case the sentiment associated with a given snippet and the association of individual texts with locations.

The first of these steps, preprocessing, is more or less independent of the analysis which follows (though the choices made here may influence our results). Typical pipelines initially chunk corpora into documents, and then divide documents (in this case, individual accounts from the CLDW or photographs from Geograph) into sentences and tokens (tokens are the finest units of analysis with which we concern ourselves, and can include words, punctuation and other processed elements of text such as lemmas or stems (Manning and Schutze 1999)). In the case of Geograph, extracting sentences is a relatively simple task, since the texts largely consist of short captions for the associated image. The CLDW, on the other hand, presented more challenges, including idiosyncratic punctuation, case and hyphenation (Butler *et al*. 2017, Donaldson *et al*. 2017). Nonetheless, in both cases sentence extraction and tokenisation were carried out using the NLTK Python Library with no modifications.[5] Having tokenised the texts, we carried out part of speech tagging, removed stop words and normalised all tokens to lower case.

4.2. *Seeking the silence*

An initial reading of our two corpora showed that both contained a wealth of descriptions relating to peaceful sounds in the Lake District, and our first task was to build two sub-corpora containing only text snippets that related to such acoustic experiences. As we have seen, over the course of the period in which we are interested here, interpretations of quietness shifted; it transitioned from being understood as a symptom of a lack of civilisation into a desirable characteristic for the experience of tranquillity. In light of this change, we used the Historical Thesaurus of English to generate a set of seed terms that are related to quiet sounds in the 'Inaudibility', 'Faintness/weakness' and 'Quietness/Tranquillity' categories. We removed terms which were only in use before 1750, but retained those which were either current in the eighteenth or nineteenth centuries, or which came into use prior to that date and have continued to be used in the present day. An initial search demonstrated that a subset of the terms thus selected were highly ambiguous in both the CLDW and Geograph extracts (e.g. *rest, sleep, dead*) and we removed these from the list. To further minimise the effects of word sense ambiguity, we used part of speech tagging to filter more terms; for instance, we retained *still* when used as an adjective (sense: 'not moving or not making a sound'), but removed it where it was employed as an adverb (sense: 'even now'). It is important to note that

our aim in these searches was to maximise the recall of descriptions related to silence, while achieving reasonable precision.

The resulting sub-corpora consisted of sentences which were likely—though not guaranteed—to be related to quiet sounds or peaceful experiences. Table 1 shows some of these sub-corpora's basic properties in relation to the main corpora from which they were extracted. Interestingly, potential descriptions of silence are more or less ubiquitous in the CLDW (89% of authors referred to silence), but much rarer overall in Geograph (mentioned by only around 10% of authors). This difference might illustrate the dominance of the visual in Geograph (since the descriptions relate to photographs), but also implies the importance of quietness for historical interpretations of the Lake District; as the author Frederick Amadeus Malleson put it, 'external nature, with all her charms, can only occupy the mind in its leisure hours of quiet peace and meditation'. In other words, writers like Malleson discovered that they could only connect with the Lake District's beauties when they could enjoy the scenery in tranquillity.

The texts from the CLDW also contain a much more varied vocabulary than the Geograph source; more than this, comparing the use of parts of speech in the sub-corpora with the main corpora reveals that the use of all parts of speech, both in quantity and range, is statistically significantly higher in the CLDW (randomisation test, $p < 0.005$). While we did find similar numbers of relevant sentences in both corpora, with similarly rich vocabularies in terms of the number of nouns and adjectives used, the contemporary descriptions nevertheless use significantly fewer unique nouns (815 vs. 2434) and adjectives (291 vs. 922) than the historical sub-corpus in their descriptions of silence. Meanwhile, there are fewer and less varied nouns in the Geograph sub-corpus, and the quantity—though not the range—of adjectives is significantly more in our selection than in the corpus as a whole (randomisation test, $p < 0.005$). In part, the CLDW's greater linguistic variation may be attributable to the fact that the texts from this corpus tend to have longer sentences, but this difference alone seems unlikely to account for the magnitude of the change. Paying closer attention to specific types of description, as well as to individual accounts, may reveal why these changes occurred, and indicate shifts in the perception of Lakeland sounds and silences.

4.3. *Unpacking the experience of silence*

Having established that our search terms retrieved rich descriptions, we started to explore what these accounts revealed about changes to experiences of the Lakeland soundscape from the eighteenth century to the present day. We began by looking at the relationship between our search terms and the sub-corpora. Of the 66 seed terms with

Table 1. Comparison of the properties of the corpora and sub-corpora.

Corpus	CLDW		Geograph	
Version	Full corpus	Extracted silence	Full corpus	Extracted silence
Number of unique authors	55	49	1076	118
Number of sentences	43,269	590	98,206	362
Mean length of sentences (with/without punctuation)	30/26	47/41	14/13	18/16
Number of nouns (total/unique)	306,722/25,057	6271/2434	417,455/21,956	1730/815
Number of adjectives (total/unique)	102,530/9759	2595/922	122,670/10,517	848/291

which we searched, 28 returned sentences in the CLDW and 14 in Geograph. Furthermore, in Geograph only 6 terms occur more than 5 times (*quiet, peaceful, calm, peace, tranquil* and *quietly*), while in the CLDW 17 terms had a frequency greater than 5. In both corpora, *quiet* was the most prevalent term; it featured in 60% of extracted sentences in Geograph, and in 16% of sentences in the CLDW sub-corpus. *Silence* presents a more complex example. It is relatively common in the CLDW and was the 3rd most frequent search term, returned in 11% of sentences. However, in the Geograph sub-corpus, it was very rare (we found only a single occurrence). Why these differences exist requires further unpacking.

To develop a more nuanced understanding of changes in the interpretation of these terms, we began by exploring co-occurrences between our seed terms and other terms found in the extracted sentences. We firstly calculated all co-occurrences at the sentence level after removing stop words. 1904 terms co-occurred more than twice in the CLDW, while only 407 did so in Geograph, reflecting the different distributions of unique parts of speech found in Table 1. To tease out the semantics at a macro level in our texts, we identified four commonly occurring classes in the top 100 co-occurrences: references to **natural** or **anthropogenic** objects (e.g. *lake, mountain, house, road*), references to **time** (e.g. *morning, instant, time, afternoon*) and references to **generic locations** (e.g. *scene, view, spot*). In Geograph, 63 of the top 100 co-occurrences could be allocated to these classes; in the CLDW, the same was true of 51 of the top 100. We used these classes, the most common co-occurences of which are shown in Table 2, as the basis for the annotation we describe below. Many of the remaining terms described properties of objects (e.g. *old, green, little, deep*), emotions (e.g. *happy, love*) or spatial prepositions (e.g. *near, distant, close*).

Natural features were the most common class in both sub-corpora, although this is to be expected since some of these terms (e.g. *dales, tarn* and *fells* in Geograph and *lake* and *sky* in the CLDW) occur equally or more often in the corpora as a whole. Anthropogenic terms associated with sound are dominated in Geograph by *road*, and in particular the *M6*, a *motorway* which runs along the edge of the National Park. In the CLDW the significant co-occurrences also relate to transport, but suggest movement by

Table 2. Classified co-occurrences. Words denoted with an asterisk occur significantly more often in the silence sub-corpora than in the random subsets of the corpora (randomisation test, $p < 0.005$). Multiple values (e.g. 11 +10 + 9) denote co-occurrences with more than one seed-term.

	CLDW		Geograph		CLDW		Geograph	
Corpus	word	count	word	count	word	count	word	count
Class	Natural				Anthropogenic			
	nature*	11	lake*	15	house	8	road*	24
	lake	11 + 10 + 9	valley*	15	man*	7	park	17
	vale*	9	dales*	9	boat*	7	lane	10
	clouds*	8	tarn	8	gentlemen	6	M6*	10
	sky	7	fells	8	town	6	motorway*	9
Class	Generic locations				Time			
	way	9	spot*	14	day*	12 + 8	day*	12 + 10
	country	8	place*	14	evening*	7	morning*	10
	scene*	8 + 7	district*	13	time	6 + 6	early*	8
	spot	7	corner*	12	instant*	6	times*	8 + 8
	paradise	7	area	10	morning*	6	Sunday*	7

boat, and also indicate human habitation. By contrast, generic place descriptions (e.g. *spot, place* and *scene*) are often used to characterise locations which are discussed with respect to sound. These descriptions seem to imply not only aural, but visually perceptible locations (e.g. a quiet spot or a peaceful valley). In Table 2 we do not, for reasons of space, show the seed terms with which co-occurrences occurred. For Geograph, these are dominated by *quiet*. Other terms, such as *calm* and *peace*, are also prominent in the CLDW. *Calm* in particular is often found in conjunction with weather and water-related terms. As we will see in more detail shortly, calmness implies not only an absence of noise, but also movement.

We have reached, for now, the limits of what can be achieved by slicing and dicing our corpora and, guided by these observations, now resort to alternative methods to analyse the nature of the references to silence and tranquillity in the CLDW and Geograph. As we will now explain, in order to annotate the two sub-corpora, we combined the results from this macroanalysis with a micro-analytic approach adopted from literary studies. In this way, our annotation agreement is perhaps the most cohesive evidence of a middle ground practice which establishes a dataset that is especially suited for the kind of multiscalar, multidisciplinary textual analysis for which we advocate here.

4.4. *Characterising the silence*

To better understand the nature of our two corpora, we developed a two-layer annotation scheme. The first layer of annotation aimed to capture the nature of the sounds and sound-related descriptions found in our texts. Based on our macroreading, we proposed the following broad classifications:

- **Total silence and tranquil sounds**: Either explicit descriptions of complete silence or a combination of tranquil sounds without contrast (e.g. *the silence was total*).
- **Contrasting sounds**: Descriptions capturing ephemerality in silence or tranquillity at a location, comparing one location to another that is less tranquil, or mentioning a sound which adds (or detracts) from the overall tranquillity (e.g. *we heard nothing but the hum of the bees*).
- **Combination of visual and aural**: Silence is implicit in the overall description of a scene, and visual properties are also conveyed (e.g. *a quiet spot above the lake*).
- **No movement**: Implied silence or tranquillity, but explicit mention of a lack of movement (e.g. *yachts sit at anchor in this quiet bay*).
- **Not relevant**: Search terms used in another sense, sounds which do not convey silence or tranquillity or descriptions of sounds not situated in the landscape (e.g. *the clock ticked loudly*).

The second layer of our annotation scheme relates to the nature of the potential sound emitters in a description. Here, we follow Krause (2008) in that, where an explicit mention of a potential sound emitter was made, we allocated it to one, or a combination thereof, of the following classes:

- **Geophony**: Natural sounds produced by non-biological sources (e.g. *wind, thunder, waterfalls*).

- **Biophony**: Natural sounds produced by animals (e.g. *lowing cows* or *humming bees*).
- **Anthrophony**: Sounds produced by humans either directly or indirectly (e.g. *noisy kids* or *busy road*).

Annotation of text is often challenging, and the texts in the CLDW were particularly difficult to interpret. To mitigate the texts' ambiguities as much as possible, we carried out an iterative annotation process. Two of the authors annotated 10% of each sub-corpora, discussing disagreements and refining unclear guidelines. After three iterations (i.e. annotating 30% of both sub-corpora), an inter-annotator agreement (Cohens Kappa) of 0.88 for types of silence and 0.90 for sound emitters in the Geograph corpus was reached. According to Landis and Koch (1977) this level of agreement is 'almost perfect', and a single annotator then annotated the remaining 70% of Geograph texts. For the CLDW, after three rounds of iteration we plateaued at 'substantial' inter-annotator agreement of 0.62 (type of silence) and 0.6 (sound emitter) respectively. Both annotators therefore annotated the remaining 70% of CLDW texts, and for cases where annotations differed discussed the texts until we reached a consensus. Table 3 shows the first layer of our annotation as absolute counts.

Three main characteristics are striking when the texts are processed in this way. First is the almost complete absence of total silence and tranquil sounds in Geograph, suggesting—as we observed above—that the lack of descriptions using the search term *silence* was indeed indicative of a change in the way the Lake District soundscape is perceived. Second is the much larger proportion of extracted descriptions found in the CLDW which were not relevant for an inquiry into acoustic experiences. This is despite our use of a historical thesaurus, which we expected to be more effective at extracting descriptions from the CLDW than Geograph. There are, we think, two reasons for this result. Firstly, Geograph descriptions are less complex and more literal. Secondly, there is a demonstrably diachronic variation in language, illustrated by the ambiguity of our search terms with respect to the CLDW (e.g. *quiet*: 37%, *peace*: 69%, *quietly*: 72%). This variation once more highlights the importance of our interdisciplinary approach, and the importance of a microreading of the texts.

The third key characteristic of this first layer of annotation concerns the overall distribution of classified sounds. The overall ranking, if not the proportion, is the same for both sub-corpora; descriptions of a combination of visual and aural are most common, followed in decreasing rank by contrasting sounds, no movement, total silence and tranquil sounds.

Figure 2 illustrates the distribution of potential sound emitters as a function of our sound classes. Biophony is relatively uncommon in both sub-corpora, whereas the presence of geophony in all the classes demonstrates the importance of the physical

Table 3. Counts of descriptions per class.

Corpus/Class	CLDW	Geograph
Total silence and tranquil sounds	46	3
Contrasting sounds	70	108
Combination of visual and aural	168	179
No movement	48	40
Not relevant	258	32

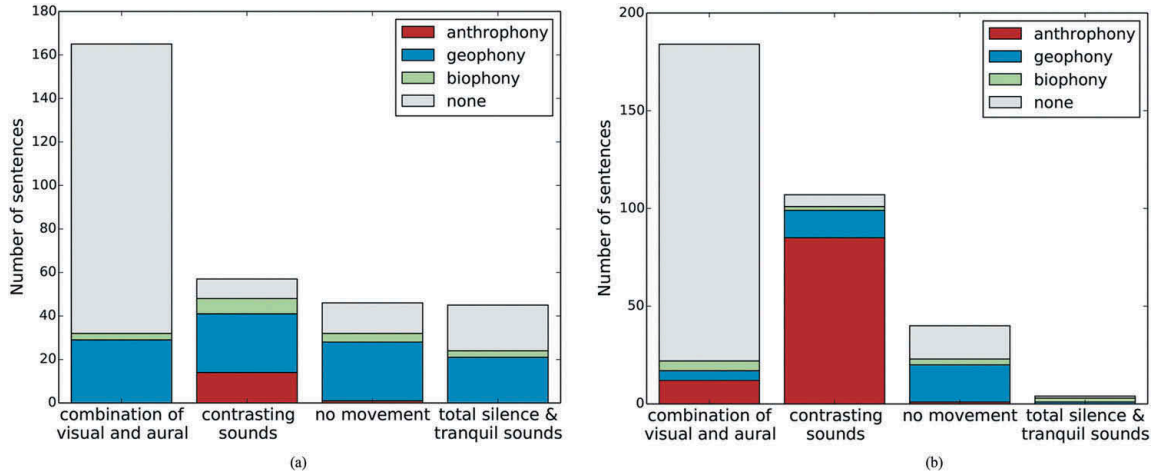


Figure 2. Number of sentences per silence class grouped by sound emitters: (a) CLDW (b) Geograph.

landscape and weather as perceived emitters of tranquil sounds in the Lake District. Yet the almost complete absence of anthrophony in the CLDW, and its dominance in the Geograph 'contrast class', stands out. This difference, we suggest, can be traced back to William Wordsworth, who constructed a certain type of cultural soundscape that affirmed the Lake District's status as a region of peace and tranquillity which was increasingly unusual in industrialising and urbanising Britain (Taylor 2018). Wordsworth deliberately overlooked noises that contrasted with the tranquil cultural soundscape he preferred; the constant cacophony from the mining industry that proliferated in the Western Lakes, like the quarries and other sites of industrial development that could be found throughout the region, are 'artfully' ignored in his works (Levinson 1986). Instead, Wordsworth's writing encouraged generations of writing about and visitors to the Lake District to focus on the sounds of flowing water and breezes blowing through the leaves on trees (our geophony), rather than the more realistic—but less pastoral—sounds of industrial activity. In Geograph, by contrast, tranquillity is valued not by ignoring the intrusion of anthropogenic influences, but rather explicitly contrasting tranquil locations to anthrophonic intrusions.

The importance of geophony, rather than anthrophony, in the CLDW is emphasised by the word clouds we created to show more clearly the patterns in the class of contrasting sounds in Figure 3. The word clouds showed the 150 most frequent nouns in the two sub-corpora, and the importance of geophony in the CLDW is underlined through the prominence of terms such as *thunder, cataracts* and *waterfalls*. A previously documented interest in echoes in this period is also visible

Figure 3. 150 most frequent nouns extracted from the class 'contrast': (a) CLDW (b) Geograph.

(Taylor 2018). In Geograph, the prominence of transport as a potential sound emitter is again clear (e.g. *road, M6, motorway* and *car*), while there is also evidence of the presence of other people as a source of discordant sounds (e.g. *visitors* and *walkers*).

The most significant finding from our macroanalysis and the resultant annotation is the revelation that what is meant by quietness has undergone a significant shift in the intervening years between the two corpora. How that features in our texts, and why that might be the case, requires more detailed focus on key individual texts within the wider collective on which we have focused so far.

4.5. *Reading the silence*

Once we had established these annotation rules, we applied them in order to ascertain quantitatively whether these terms possessed positive or negative connotations in our sub-corpora. We assigned mean sentiment values to each description using an existing Opinion Lexicon (Hu and Liu 2004) and a pretrained set of GloVe word embeddings (Pennington *et al*. 2014) to attach sentiment values to words not contained in the lexicon. Figure 4 shows histograms of sentiment for randomly selected sentences from both corpora, as well as our two sub-corpora. As the bias towards the right in the histograms indicates, descriptions of quietness tend to be positive. Secondly, both of our sub-corpora are statistically significantly more positive (t-test, $p < 0.005$) than the corpora from which they are extracted. This difference is much more marked for Geograph (overall corpus mean sentiment 0.80 ± 0.63 vs. silence sub-corpus mean sentiment 1.90 ± 0.59) than in the CLDW (overall corpus mean sentiment 1.58 ± 0.59 vs. silence sub-corpus mean sentiment 1.80 ± 0.49).

To test if this bias towards positive values can be explained by the presence of our seed words in the descriptions of silence, we calculated the sentiment values for descriptions without taking into account our seed words. The absolute mean value of the difference between the original sentiments and sentiments without seed words is small (0.4 in Geograph and 0.2 in the CLDW). Therefore, we concluded that such descriptions are in general associated with positive sentiment.

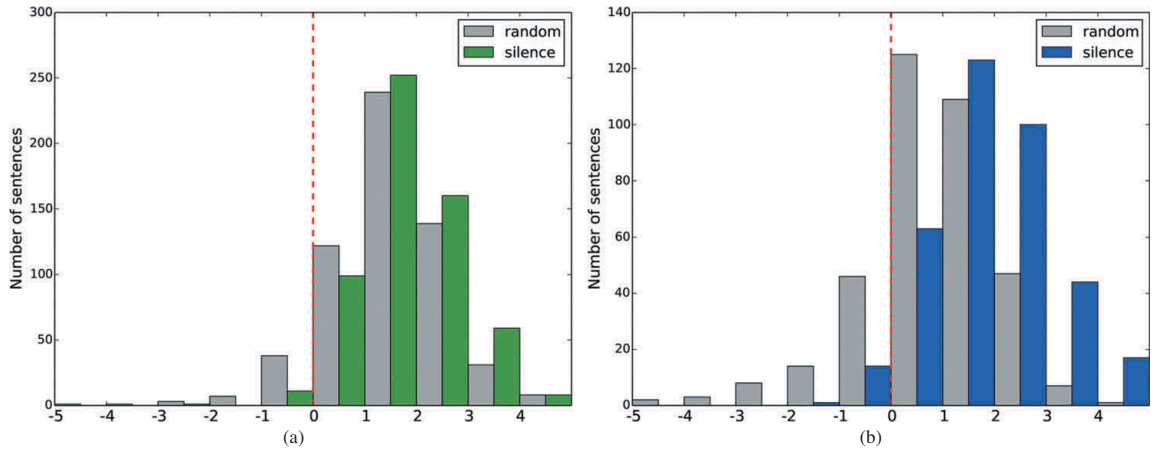


Figure 4. Number of sentences grouped by sentiment values and compared against random sample (a) CLDW $n_{sentences} = 590$, t-test $p < 0.005$, (b) Geograph $n_{sentences} = 362$, t-test $p < 0.005$.

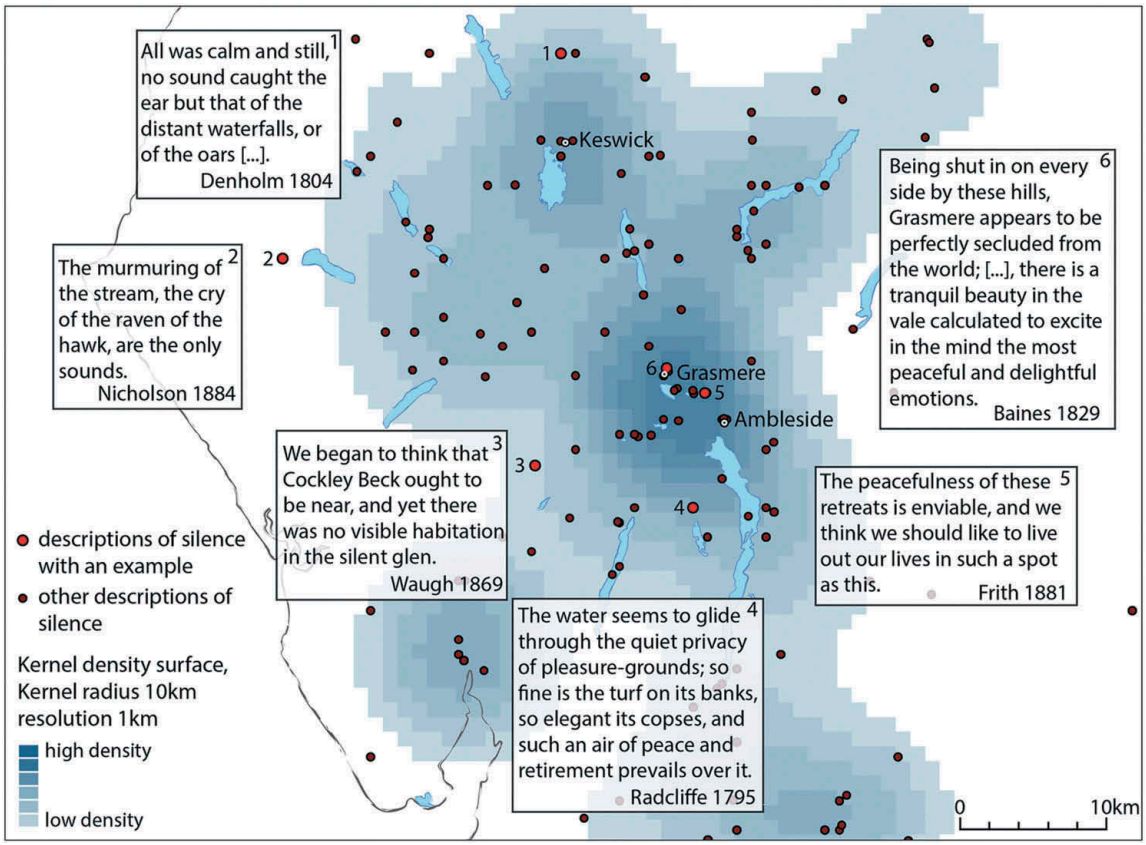


Figure 5. Locations of texts in the Lake District associated with toponyms in the CLDW and an associated kernel density surface.

The advantages—as well as the limits—of our approach in general are perhaps best illustrated by this sentiment analysis. We find that our text snippets are generally positive, and descriptions related to silence in Geograph are more positive than the corpus as a whole. The assumption—made by us and by the Opinion Lexicon—that quietness and peace relate to tranquillity, and that these are desirable qualities in a rural landscape, seems common-sensical today (MacFarlane *et al.* 2004). Evaluating why this is the case, though, relies on closer attention to individual accounts within the corpora.

We saw earlier that it has certainly not always been the case that quietness was a desirable feature (Fisher 1999), and the greater linguistic variety of the CLDW indicates that quietness was both a more common and a more complex phenomenon. William Gilpin, the Cumbrian curate most famous for his development of the picturesque mode of landscape evaluation, is influential over the promotion of quietness in the Lake District (Taylor 2018). In his *Observations, Relative Chiefly to Picturesque Beauty* (1786), Gilpin wrote of Lorton vale in the north-western Lakes that it was a place that could 'pretend not to dignity'; it could only aspire to be a 'mere [scene] of tranquillity'. Nevertheless, such a place held its own charms for Gilpin, not least because they had the potential to transport him into a particular mental state. He continued that he 'might have wished for a quiet, tranquil hour, when the glimmering surfaces of things are sometimes perhaps more pleasing – at all times certainly more soothing, than images of the brightest hue'. Gilpin grammatically links *quiet* with *tranquil* here. Indeed, the lack of sound and movement in

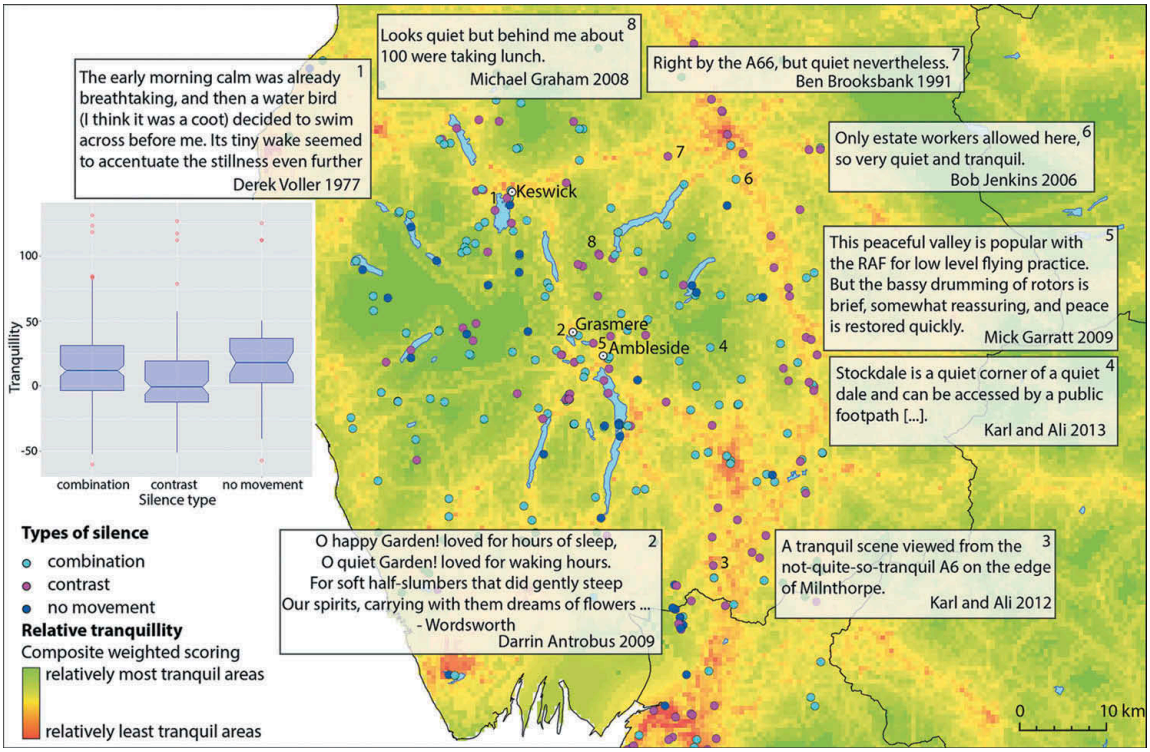


Figure 6. Comparison of the map of relative tranquillity MacFarlane 2004 and types of silence extracted from the Geograph corpus. National Tranquillity mapping data 2007 developed for the Campaign to protect rural England and Natural England by Northumbria University. OS Licence number 100018881.

the surrounding area seems to calm his mind in ways that allow him to connect more effectively with the surrounding landscape; it is only in this quiet and tranquil state that Gilpin can appreciate the 'glimmering surfaces of things' around him.

This sense of quietness as indicating acoustic peace, physical stillness and—crucially—a closely related sense of mental calm was inherited by later Lakeland authors. Joseph Mawman, writing a couple of decades after Gilpin, discovered a similar 'harmony with the soothing quiet which prevailed all around'. As for Gilpin, this quietness established for Mawman a sense of what we might recognise today as mindfulness: the quiet 'disposed us,' he wrote, 'to reflect seriously upon that interminable question, 'What state of life is best fitted for happiness?' Later still, the Lancashire poet Edwin Waugh was even more explicit: he thought that '[g]oing from a crowded city into this little monastic town [Cartmel] is almost like going to bed, or sinking into an antiquarian dream, – all is so quaint and quiet'. For Waugh, quietness does not simply establish a calm and meditative state; it actually feels like going back in time to an earlier moment when the world was, he imagines, less frantic.

This meaning of *quiet* as indicating peacefulness in the landscape as well as a form of mental calm is not remarked on by the Geograph writers. That is not, however, to say that the connection does not exist. Karl and Ali, for instance, describe Stockdale as 'a quiet corner of a quiet dale', and here the repetition of the search term highlights the writers' enhanced sense of being removed from the busier, more inhabited parts of the region, and emphasises a certain stasis in the valley. Bob Jenkins, meanwhile, implies a Gilpinian quietness when he writes of Lowther Park that since 'only estate workers

[are] allowed', it is 'very quiet and tranquil' (Figure 6). While in the CLDW, the writers needed to explain this link between the physical soundscape and their own mental state, it is taken for granted in the Geograph texts. These later authors therefore require less diverse language to indicate the sense of tranquillity they discover in the Lake District's quiet places; they, like us, have inherited from writers such as those represented in the CLDW a sense that quietness is a positive value. The difference is that, for the Geograph authors, quietness is a much scarcer commodity than it was for the CLDW writers (even at a historical moment that, elsewhere, was witnessing an exponential increase in noise (Picker 2003, Bour 2016)). Nevertheless, as Timothy Morton writes, 'in order to differentiate [quietness] there must be some roughness, some *noise*' (2009, p. 71), and the writers in both our corpora share a sense of relief at discovering spaces of quiet in an increasingly noisy world.

4.6. *Mapping the silence*

In a final exploration of our corpora, we set out to discover not only what was talked about, but where. Place names in the CLDW have been georeferenced using toponym recognition and resolution (Rayson *et al*. 2017) and used to explore broader themes, including local variation in the use of aesthetic language (Donaldson *et al*. 2017) and acoustic experience (Taylor *et al*. 2018). In contrast, the Geograph corpus explicitly links descriptions to 1km grid squares, and texts can therefore be mapped without any additional processing. Since the two corpora document space in different ways, and at very different scales, the spatial extents which can be associated with texts are not directly comparable. In the following, we seek to once more find a middle ground, and map the two corpora to space in ways which are appropriate given these considerations.

The toponyms in the CLDW situate the texts in space, but it is often not possible to assign coordinates to the areas they describe with much accuracy. Associating extracted sentences and related sounds with locations requires us not only to identify corresponding toponyms, but also to specify an appropriate document scope that relates the content to a location (Andogah *et al*. 2012). However, the complexity of the sentence structures in the CLDW frequently makes this process challenging. For example, when James Denholm describes movement on Derwentwater—the lake beside the town of Keswick—the closest toponym actually refers to a nearby mountain summit: 'the hills upon the left were in the shade, as was the mountain of Skiddaw, lying, together with the islands, directly in front. All was calm and still, no sound caught the ear but that of the distant waterfalls, or of the oars, striking, in alternate succession […] the surface of the lake'. Even where we can identify a named feature, associating it with a spatial extent is still challenging. On what part of Derwentwater did Denholm row, for instance, and at what point on his journey did he reflect on that state of calmness? Nevertheless, these toponyms do allow us to identify the area being discussed, even if we often cannot map it with any great degree of specificity.

In light of this complexity, we used a simple heuristic to associate toponyms with sentences. We first identified any toponyms in the target sentence; if none were found, we looked firstly one sentence back, and then one sentence forward in order to identify the most appropriate document scope. Where more than one toponym was found, the sentence was associated with multiple locations. For the 332 descriptions of silence

extracted from the CLDW, we could map a little less than half (157) to locations in this way. Given the above limitations, we used a kernel density estimation with a coarse kernel of 10km that reflects the uncertainty registered in the texts, but still allows us to explore spatial patterns in the data (Figure 5).

We can see from Figure 5 that Grasmere emerges in this kind of analysis as being particularly significant for discussions of sound in the CLDW. Grasmere's importance is, in some ways, not surprising: it was located on the main route between Ambleside and Keswick, and as a result was an almost inevitable sight on any Lakeland tour in the period (Murrieta-Flores *et al*. 2017). Yet, the texts in the CLDW point to another reason why Grasmere was considered to be a particularly evocative location for the pursuit of quiet repose. The poet William Wordsworth (1770–1850) resided in or near the village from 1799 until his death, and he celebrated the peace he enjoyed there in his writing. More than that, though, he also highlighted certain acoustic experiences: the echo of his sister-in-law's laugh around the valley, for instance, encouraged several imitators (Taylor 2018). Loughrigg Tarn also became a popular excursion for tourists in pursuit of the sounds of Wordsworthshire (Donaldson *et al*. 2015).

Wordsworth's influence can, in fact, be traced through to the modern day: one of the contemporary descriptions of Grasmere quotes the poet verbatim (Figure 6) to evoke the peace that allowed Wordsworth—and, perhaps, his modern reader—to enjoy 'soft half-slumbers' in the tranquil valley. The greater degree of precision offered by, and the contemporary relevance of, the Geograph corpus meant that we were able to compare this data directly with MacFarlane's *et al*.'s (2004) tranquillity study. To do so, we resampled the 500m *map of relative tranquillity* to the 1km resolution of Geograph using bilinear interpolation. Figure 6 shows the locations of the three most prominent types of silence found in these data (c.f. Table 3) along with box plots of tranquillity values. From this, it seems that contrasting sounds are associated with low values of tranquillity, particularly near the M6. The emergence of this area indicates that places associated with silence and tranquillity have 'spread' from the central Lake District that had been the focus for nineteenth-century travellers. Instead, today's visitors find that almost the entire National Park offers a sense of tranquil quiet.

A similar outcome occurs when we analyse this corpus quantitatively. Using a non-parametric Kruskal-Wallis test, we analysed tranquillity as a function of silence. A significance level of $p < 0.01$ was used to reject the null hypothesis that all types of silence were associated with similarly distributed values of tranquillity. Since this test was significant, we used a post-hoc Dunn test to compare all tranquillity values with a significance level of $p < 0.01$[6] (Table 4) (Dunn 1961).

These tests revealed that contrasting sounds in the Geograph texts are statistically significantly associated with lower values of relative tranquillity than both combination and no movement, based on an independently created model (MacFarlane *et al*. 2004). This link suggests that, firstly, the prominence of anthrophony in contrasting sounds (Figure 2) reflects real variation in environmental properties, since low values of tranquillity are typically associated with anthropogenic disturbance. Secondly, the comparison demonstrates that the texts we extracted are in broad agreement with an independently produced model. Thirdly, it also demonstrates a significant strength of our multidisciplinary approach: this method has allowed us to identify locations which

Table 4. Adjusted significance values for a post-hoc Dunn test comparing distributions of tranquillity values associated with types of silence.

Comparison	*p*.adj
Contrasting sounds vs. Combination of visual and aural	0.003
No movement vs. Combination of visual and aural	0.247
Contrasting sounds vs. No movement	0.002

are considered tranquil despite an unpromising setting. For example, Ben Brooksbank's describes a scene '[r]ight by the A66' that is 'quiet nevertheless.' Brooksbank is indicative of a significant group of authors in the Geograph corpus who identify locations beside busy roads as being comparatively peaceful. It seems that, for the modern visitor, complete quiet is not necessary for the discovery of tranquillity.

5. Discussion

Our aim in this project was to extract and interpret descriptions of, and diachronic and spatial variation in, perceived silence from historical and contemporary textual descriptions. By adopting a blend of methods, focussing on detailed reading of individual texts, annotation and stratification of descriptions of silence and a range of quantitative analyses of both corpora we were able to fulfil this aim.

Although both corpora contained references to our silence-related seed words, these were much more prominently used by authors in the CLDW (89%) than Geograph (10%). This, we argue, reflects the importance of peace and silence in historical accounts of the Lake District. Descriptions of silence were also, in the historic texts, associated with a richer use of both nouns and adjectives. This reflects, on the one hand, the more literal nature of the short descriptions in Geograph and, on the other, the need to set out the authors' mental state in a description of silence or peace in the CLDW.

To find silence-related descriptions we used a variety of seed terms. These also demonstrated a clear diachronic change, with descriptions of total silence or calmness almost totally disappearing in Geograph. Exploring terms which co-occurred with our seed terms helped us identify changes in the nature of terms associated with silence. Here we see two changes over time. Firstly, nouns associated with transport (e.g. *road, motorway*) emerge as common co-occurrences in Geograph descriptions of silence. Secondly, we note that generic place descriptions (e.g. *spot, place, corner*) become increasingly important, reflecting perhaps a change from the description of a whole landscape, to a specific location within it.

These first explorations of our corpora guided the following classifications of both silence and related sound emitters. Having identified four key classes of sounds, we annotated the extracted descriptions. This annotation further demonstrated the almost complete absence of total silence and tranquil sounds in the contemporary data, and also showed the increased importance of silence expressed through contrast. By annotating sound emitters, we identified the concern about anthropogenic disturbances in the modern landscape. Both corpora privilege descriptions of geophony over biophony, and in doing so adhere to a version of the cultural soundscape that can be traced back to writers like Wordsworth.

Wordsworth's influence on historical descriptions of the Lake District is clearly visible in the general positivity associated with descriptions of this landscape and silence within

it (Figure 4). By contrast, Geograph descriptions are in general neutral, reflecting the aim of the collection to describe the landscape. Nonetheless, descriptions referring to silence are statistically significantly more positive, reflecting value given to silence as a cultural resource. By projecting our descriptions into space, the persistent influence of Wordsworth is emphasised. In both corpora, we find a cluster of descriptions centred around Grasmere, a location popularised and written about by Wordsworth and his followers. Comparing contemporary descriptions of silence to a map of relative tranquillity showed that contrast is both semantically and spatially associated with anthropogenic disturbance. This comparison also illustrates how our textual descriptions can indeed allow us to identify tranquil locations even in busy areas of the landscape.

Our aim in this work was to uncover a middle ground that combines interdisciplinary methods to generate multiscalar perspectives on textual, spatial data. Pragmatically, if we wish to make a contribution to Landscape Character Assessment, this result matters since it demonstrates two key points. Firstly, the prominence of descriptions in our contemporary corpus which refer to generic places (e.g. *spot, place, corner*) implies a form of landscape perception that focusses on locations with some form of gestalt coherence (Schroeder 2007). Secondly, modelling relative tranquillity is contrary to current GIS-based attempts at quantifying such properties, which often focus on distance from potential emitters as a proxy for disturbance (e.g. Carver *et al*. 2002, MacFarlane *et al*. 2004). Rather, our approach suggests an additional need for modelling tranquil places by contrast, as suggested—though in a very different context—by Winter and Freksa (2012). Further, this approach points to an oft-observed dichotomy between attempts to model landscape properties as continuous fields (Mücher *et al*. 2010) and the diverse ways in which people perceive and categorise the world (Mark *et al*. 2011).

It is, of course, important to note a number of limitations with our approach. Firstly, our results are dependent on the choices we made during preprocessing, including: the seed words selected; the reliability of our annotation; and the specific methods we used (e.g. the quality of the part of speech tagging, the use of GloVe embeddings and our approach to sentiment analysis). However, though such limitations are part and parcel of any text-based approach, we argue that our results are robust since quantitative macroreadings of our corpora were interpreted through, and substantiated by, qualitative microreadings. Secondly, our corpora have different properties, particularly with respect to georeferencing and granularity. Putting aside the inevitable uncertainty introduced by mapping toponyms directly to point locations, the rich descriptions found in the CLDW cannot be easily mapped to areas associated with places described in the texts (Murrieta-Flores *et al*. 2017). We suggest that until methods such as those proposed by Moncla *et al*. (2016) can be applied successfully to historical texts, spatial comparisons of this kind are best performed on the region as a whole (c.f. Figure 2).

6. Conclusions

We set out to explore how finding the middle ground—a place for a blend of methods from a range of disciplines—could offer us insights into two temporally distinct, spatially overlapping corpora describing experiences of the Lake District landscape. In particular:

- Unstructured texts offer rich, semantically diverse, and spatially groundable insights into landscape perception, and more generally access to understandings of the way place is made and conceptualised.
- Diachronic use of corpora offer insights into ways in which readings of contemporary and historical landscape descriptions are intertwined.
- Spatial contiguous models of properties such as tranquillity can be enhanced and refined through complementary analysis of spatially grounded textual sources.

Our results do not necessarily suggest new ways of understanding silence and soundscapes. Rather, they reveal scalable approaches towards exploring how people represent, in writing, their individual experiences of landscapes in given places. Practically speaking, our approach suggests ways of extracting and analysing important information required in Landscape Character Assessment, and could be scaled up to cover large spatial extents. More generally, we suggest that GIScience would do well to consider the opportunities offered by critically exploring rich unstructured text, whilst literary historical studies should embrace the plethora of authors and viewpoints offered by this kind of approach. For both disciplines, this middle ground offers a way of increasing the breadth of participation in the production of spatial information and knowledge.

Notes

1. https://www.nps.gov/subjects/sound/soundsmatter.htm.
2. https://github.com/UCREL/LakeDistrictCorpus.
3. http://www.geograph.org.uk/.
4. Based on an anonymous survey carried out by the initiators of the project.
5. https://www.nltk.org/.
6. p value adjusted for multiple means using the Benjamini-Yekuteili method https://www.rdocumentation.org/packages/FSA/versions/0.8.20/topics/dunnTest.

Acknowledgments

We would like to gratefully acknowledge all the contributors to Geograph British Isles (Creative Commons Attribution-ShareAlike 2.5 License). We are grateful to Graeme Willis (Campaign to Protect Rural England) and Nick Groome (Ordnance Survey) for their help in accessing the National Tranquillity Mapping Data. Finally, we thank the reviewers for their useful suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Research for this article was partly supported by the Leverhulme Trust-funded project 'Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District' (RPG-2015-230). IGs contribution to this paper received funding from the European Research Council (ERC) under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850). RSP gratefully acknowledges support from the

Swiss National Science Foundation Project EVA (166788) and the University Research Priority Programme on Language and Space.

Notes on contributors

Olga Chesnokova is a PhD student at the University of Zurich. Her focus lies in the use of unstructured text to better understand landscape and landscape perception.

Joanna E. Taylor is Presidential Academic Fellow in Digital Humanities at the University of Manchester. Her research uses digital humanities methodologies – particularly digital mapping – to explore literary geographies and environmental histories from the long nineteenth-century.

Ian N. Gregory is Professor of Digital Humanities at Lancaster University. His research interests lie in the Spaital Humanities, and particularly how GISc techniques can be used in combination with other digital humanities approaches, to analyse large bodies of text.

Ross S. Purves is a Professor at the Department of Geography of the University of Zurich. His research interests focus on how we can answer and explore societally relevant geographic questions paying attention to vagueness and uncertainty, often using unstructured data in the form of text as a primary source.

ORCID

Olga Chesnokova http://orcid.org/0000-0002-4298-1789
Joanna E. Taylor http://orcid.org/0000-0001-8597-0097

References

Agnew, V., 2012. Hearing things: music and sounds the traveller heard and didn't hear on the grand tour. *Cultural Studies Review*, 18 (3), 67–84. doi:10.5130/csr.v18i3.2855

Ammer, C., 2013. *The American heritage dictionary of idioms*. 2nd ed. Boston, MA: Houghton Mifflin.

Andogah, G., Bouma, G., and Nerbonne, J., 2012. Every document has a geographical scope. *Data & Knowledge Engineering*, 81–82, 1–20. doi:10.1016/j.datak.2012.07.002

Antoniou, V., Morley, J., and Haklay, M., 2010. Web 2.0 geotagged photos: assessing the spatial dimensions of the phenomenon. *Geomatica*, 64 (1), 99–110.

Bour, I., 2016. Foreword: noise and sound in the Eighteenth century. *In*: *Etudes Epistémè: revue de littérature et de civilisation (XVIe-XVIIIe siecles)*, 29. Available from: http://journals.openedition.org/episteme/11 [Accessed 15 April 2018]. doi:10.4000/episteme.1136

Burenhult, N. and Levinson, S.C., 2008. Language and landscape: a cross-linguistic perspective. *Language Sciences*, 30 (2–3), 135–150. doi:10.1016/j.langsci.2006.12.028

Butler, A., 2016. Dynamics of integrating landscape values in landscape character assessment: the hidden dominance of the objective outsider. *Landscape Research*, 41 (2), 239–252. doi:10.1080/01426397.2015.1135315

Butler, J.O., *et al*., 2017. Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Maps and Geography Libraries*, 13, 58–81. doi:10.1080/15420353.2017.1307304

Carver, S., Evans, A., and Fritz, S., 2002. Wilderness attribute mapping in the United Kingdom. *International Journal of Wilderness*, 8 (1), 24–29.

Chesnokova, O. and Purves, R.S., 2018. From image descriptions to perceived sounds and sources in landscape: analyzing aural experience through text. *Applied Geography*, 93, 103–111. doi:10.1016/j.apgeog.2018.02.014

Coates, P.A., 2005. The strange stillness of the past: toward an environmental history of sound and noise. *Environmental History*, 10 (4), 636–665. doi:10.1093/envhis/10.4.636

Comber, A., Fisher, P., and Wadsworth, R., 2005. What is land cover? *Environment and Planning B: Planning and Design*, 32 (2), 199–209. doi:10.1068/b31135

Corbin, A., 1986. *The foul and the fragrant: odor and the French social imagination*. USA: Harvard University Press.

Council of Europe, 2000. European landscape convention. *Report and Convention Florence*, ETS No. 17 (176), 8.

Donaldson, C., Gregory, I.N., and Murrieta-Flores, P., 2015. Mapping 'Wordsworthshire': a GIS study of literary tourism in Victorian Lakeland. *Journal of Victorian Culture*, 20 (3), 287–307. doi:10.1080/13555502.2015.1058089

Donaldson, C., Gregory, I.N., and Taylor, J.E., 2017. Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography*, 56, 43–60. doi:10.1016/j.jhg.2017.01.006

Dubois, P., 2016. The impossible temptation of noise in late Eighteenth-century English music. *In*: *Etudes Epistémè: revue de littérature et de civilisation (XVIe-XVIIIe siecles)*, 29. Available from: https://journals.openedition.org/episteme/1122 [Accessed 15 April 2018]. doi:10.4000/episteme.1122

Dunn, O.J., 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56 (293), 52–64. doi:10.1080/01621459.1961.10482090

Fairclough, G. and Herring, P., 2016. Lens, mirror, window: interactions between historic landscape characterisation and landscape character assessment. *Landscape Research*, 41 (2), 186–198. doi:10.1080/01426397.2015.1135318

Fisher, J.A., 1999. The value of natural sounds. *Journal of Aesthetic Education*, 33 (3), 26–42. doi:10.2307/3333700

Fisher, P., Wood, J., and Cheng, T., 2004. Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, 29 (1), 106–128. doi:10.1111/tran.2004.29.issue-1

Fuchs, R., *et al*., 2015. The potential of old maps and encyclopaedias for reconstructing historic European land cover/use change. *Applied Geography*, 59, 43–55. doi:10.1016/j.apgeog.2015.02.013

Gregory, I. and Donaldson, C., 2016. Geographical text analysis: digital cartographies of Lake District literature. *In*: D. Cooper, C. Donaldson and P. Murrieta-Flores, eds. *Literary mapping in the digital age*. London: Routledge, 67–78. ISBN 9781472441300.

Grenon, P. and Smith, B., 2004. SNAP and SPAN: towards dynamic spatial ontology. *Spatial Cognition and Computation*, 4 (1), 69–104. doi:10.1207/s15427633scc0401_5

Hegarty, P., 2007. *Noise/Music: a history*. London: Continuum International Publishing Group.

Hu, M. and Liu, B., 2004. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on knowledge discovery and data mining KDD 04*, Seattle, WA, USA, 04, 168.

Jockers, M., 2013. *Macroanalysis: digital methods and literary history*. Chicago: University of Illinois Press.

Joy, L., 2014. Relative obscurity: the emotions of words, paint and sound in Eighteenth-century literary criticism. *History of European Ideas*, 40 (5), 644–661. doi:10.1080/01916599.2013.860291

Krause, B., 2008. Anatomy of the Soundscape. *Journal of the Audio Engineering Society*, 56, 1/2.

Lake, M.W., Woodman, P.E., and Mithen, S.J., 1998. Tailoring GIS software for archaeological applications: an example concerning viewshed analysis. *Journal of Archaeological Science*, 25 (1), 27–38. doi:10.1006/jasc.1997.0197

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174. doi:10.2307/2529310

Levinson, M., 1986. *Wordsworth's great period poems: four essays*. Cambridge: Cambridge University Press.

Leyk, S., Boesch, R., and Weibel, R., 2006. Saliency and semantic processing: extracting forest cover from historical topographic maps. *Pattern Recognition*, 39 (5), 953–968. doi:10.1016/j.patcog.2005.10.018

Lowenthal, D., 1976. Tuning into the past: can we recapture the soundscapes of bygone days?. *UNESCO Courier*, November, (29), 15–20.

MacFarlane, R., *et al*., 2004. *Tranquillity mapping: developing a robust methodology for planning support*. Technical report on Research in the Northumberland National Park and the West Durham Coalfield, Northumbria University, Newcastle University.

Manning, C.D. and Schutze, H., 1999. *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Mark, D.M., *et al*., eds., 2011. *Landscape in language. Transdisciplinary perspectives*. Amsterdam/ Philadelphia: John Benjamins.

McGarigal, K. and Marks, B.J., 1994. FRAGSTATS: spatial pattern analysis program for quantifying landscapesStructure. *General Technical Report PNW-GTR-351. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station*. Portland, OR, 97331 (503), 134. doi:10.3168/jds.S0022-0302(94)77044-2

Miller, N.P., 2008. US National parks and management of park soundscapes: a review. *Applied Acoustics*, 69 (2), 77–92. doi:10.1016/j.apacoust.2007.04.008

Moncla, L., *et al*., 2016. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*, 30 (6), 1137–1160. doi:10.1080/13658816.2015.1108422

Morton, T., 2009. *Ecology without nature: rethinking environmental aesthetics*. Harvard: Harvard University Press.

Mücher, C.A., *et al*., 2010. A new European Landscape Classification (LANMAP): a transparent, flexible and user-oriented methodology to distinguish landscapes. *Ecological Indicators*, 10 (1), 87–103. doi:10.1016/j.ecolind.2009.03.018

Murrieta-Flores, P., Donaldson, C., and Gregory, I.N., 2017. GIS and literary history: advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature. *Digital Humanities Quarterly*, 11, 1.

Nomination, 2015. Nomination of the English Lake District for inscription on the world heritage list. [online]. Available from: https://whc.unesco.org/en/list/422 [April, 7, 2018].

Pennington, J., Socher, R., and Manning, C.D., 2014. GloVe: global Vectors for Word Representation. *In*: *Conference on empirical methods in natural language processing (EMNLP 2014),* Doha, Qatar.

Picker, J.M., 2003. *Victorian soundscapes*. Oxford: Oxford University Press.

Pickles, J., ed., 1995. *Ground truth: the social implications of geographic information systems*. New York: Guilford Press.

Pijanowski, B.C., *et al*., 2011. Soundscape ecology: the science of sound in the landscape. *Bioscience*, 61 (3), 203–216. doi:10.1525/bio.2011.61.3.6

Quercia, D. and Schifanella, R., 2015. Smelly maps: the digital life of urban smellscapes. *In*: *9th international AAAI conference on web and social media,* Oxford, UK.

Rayson, P., *et al*., 2017. A deeply annotated testbed for geographical text analysis: the corpus of Lake District writing. *In*: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial humanities*, Redondo Beach, CA, USA, 9–15.

Rosenfeld, S., 2011. On being heard: a case for paying attention to the historical ear. *The American Historical Review*, 116 (2), 316–334. doi:10.1086/ahr.116.2.316

Rundstrom, R.A., 1995. GIS, indigenous peoples, and epistemological diversity. *Cartography and Geographic Information Systems*, 22 (1), 45–57. doi:10.1559/152304095782540564

Schroeder, H.W., 2007. Place experience, gestalt, and the human-nature relationship. *Journal of Environmental Psychology*, 27 (4), 293–309. doi:10.1016/j.jenvp.2007.07.001

Sexton, J.O., *et al*., 2016. Conservation policy and the measurement of forests. *Nature Climate Change*, 6 (2), 192. doi:10.1038/nclimate2816

Smith, B.R., 1999. *The acoustic world of early modern England: attending to the O-Factor*. Chicago and London: The University of Chicago Press.

Smith, M.M., 2004. Introduction: onward to audible parts. *In*: M.M. Smith, ed. *Hearing history: a reader*. Athens, GA: University of Georgia Press, ix-xxii.

Smith, S.J., 1994. SoundScape. *Area,*26, 232–240.

Taylor, J., Gregory, I., and Donaldson, C., 2018. Combining close and distant reading: a multiscalar analysis of the English Lake District's historical soundscape. *International Journal of Humanities and Arts Computing*, 12 (2), 163–182. doi:10.3366/ijhac.2018.0220

Taylor, J.E., 2018. Echoes in the mountains: theromantic Lake District's soundscape. *Studies in Romanticism*, (forthcoming).

Truax, B., 1978. *Handbook for acoustic ecology*. Burnaby, B.C. Canada: ARC Publications.

Tudor, C., 2014. *An approach to landscape character assessment*. (October), 56. Worcester, UK: Natural England.

Turner, S., 2006. Historic landscape characterisation: a landscape archaeology for research, management and planning. *Landscape Research*, 31 (4), 385–398. doi:10.1080/01426390601004376

Van Den Berghe, H., *et al*., 2018. Using the past to indicate the possible presence of relics in the present-day landscape: the western front of the great war in Belgium. *Landscape Research*, 6397, 1–23. doi:10.1080/01426397.2017.1415315

Vasconcelos, M., *et al*., 2002. Land cover change in two protected areas of Guinea-Bissau (1956–1998). *Applied Geography*, 22 (2), 139–156. doi:10.1016/S0143-6228(02)00005-X

Winter, S. and Freksa, C., 2012. Approaching the notion of place by contrast. *Journal of Spatial Information Science*, 2012 (5), 31–50.

D

APPENDIX: PUBLICATION 4

Koblet, O. and Purves, R.S., 2020. From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally. Landscape and Urban Planning, Volume 197, 103757.

PhD candidate's contributions: Developing research idea, annotation, data processing and analysis, creation of maps, carrying out a face validity step with an expert group, writing the draft manuscript and incorporating co-author's feedback.

Landscape and Urban Planning 197 (2020) 103757

Contents lists available at ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally

Olga Koblet*, Ross S. Purves

Department of Geography, University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Landscape perception
LCA
Text analysis

ABSTRACT

Inspired by the narrative nature of Landscape Character Assessment (LCA), we present a complete workflow to (i) build a collection of almost 7000 online texts capturing first-person perception of the Lake District National Park in England, and (ii) analyse these for sight, sound and smell perception. We extract and classify more than 28,000 descriptions referring to sight, almost 1500 to sound and 78 to smell experiences using text analysis. The resulting descriptions can be explored for the whole Lake District revealing for example, how traffic noise intrudes on experiences in the mountains close to transportation axes. Linking descriptions to LCA areas allow us to compare properties of different regions in terms of scenicness or tranquillity at a macro-level by identifying, for example, LCA areas dominated by descriptions of tranquillity or anthropogenic sounds. At a micro-level, we can zoom in to individual descriptions and landscape elements to understand how particular places are experienced in context. Local experts gave positive feedback about the utility of such information as a monitoring tool complementary to existing approaches. Our method has potential for use both in allowing comparison over time and identifying emerging themes discussed in online texts. It provides a scalable way of collecting multiple perspectives from written text, however, more work is required to understand by whom, and why, these contributions are authored.

1. Introduction

According to the European Landscape Convention (Council of Europe, 2000) public perception should be taken into account in landscape assessment. However, in practice this is difficult (Jones and Stenseke, 2011). How do we collect the explicit opinions of people who have visited and experienced landscape? Methodologically, in-depth interviews and other qualitative approaches are one possibility, but they are typically applied only locally (Bieling, Plieninger, Pirker, & Vogl, 2014; Caspersen, 2009; Clemetsen, Krogh, & Thorén, 2011). Participatory GIS (PPGIS) and surveys are easier to use for larger areas, however, they often capture the views of local residents and exclude others interacting with landscapes (Brown & Reed, 2009; Bruns & Stemmer, 2018; Kienast, Frick, van Strien, & Hunziker, 2015). Therefore, a problem exists not only in sourcing public perception of landscapes, but also in collecting diverse voices (Butler, 2016). In this paper we combine the need to capture different groups and provide solutions suitable for large regions by collecting and computationally analysing texts describing a range of individual experiences in landscapes.

One pioneering framework in landscape assessment, initiated in the UK in the 1980s, and since adapted by many countries – Landscape Character Assessment (LCA) – aimed for a shift from describing iconic landscapes, to describing all landscapes without exception. An important goal was capturing properties distinguishing distinctive areas from their neighbours (Fairclough et al., 2018; Tudor, 2014). LCA's guidelines emphasise the importance of individual experiences in landscapes perceived through multiple senses "such as smell/scent, tranquillity, noise, and exposure to the elements (wind and rain for example)" (page 42, Tudor, 2014). The LCA process is divided into 4 steps: definition of purpose and scope, desk study, field study and final classification and description (Tudor, 2014). The desk phase collects information about physical properties of landscapes and delineates areas of distinctive character (LCA areas). Fieldwork is mostly concerned with *in situ* perception of people towards given landscapes. The results are then compiled into rich textual descriptions for each LCA area. Important challenges for LCA include integrating perspectives and perceptions from multiple people (and not only experts) and multiple senses (not overprivileging sight) (Swanwick & Fairclough, 2018). Furthermore, different groups of people value landscapes in different ways. For example, Butler (2016) adopted categories identified by Relph (1976) to a landscape context, and criticised the dominance of the 'objective outsider' in LCA. Considering other voices, and in

* Corresponding author.
E-mail addresses: olga.koblet@geo.uzh.ch (O. Koblet), ross.purves@geo.uzh.ch (R.S. Purves).

https://doi.org/10.1016/j.landurbplan.2020.103757
Received 10 July 2019; Received in revised form 7 January 2020; Accepted 19 January 2020

particular those of 'insiders' - should be part of the LCA process (Butler, 2016; Swanwick & Fairclough, 2018). These challenges are not unique to LCA and are relevant for all integrated approaches to landscape monitoring including experienced perception at some level: these may include what Kienast, Wartmann, and Hunziker (2019) term indicator-driven approaches and comprehensive narratives of landscapes.

One possible approach to addressing this gap is through the use of rich written sources, as has long been practised by environmental historians (Cronon, 1992). Galaz et al. (2010) and Daume, Albert, and von Gadow (2014) suggested using texts extracted from the internet in ecological monitoring and identifying unanticipated threats in forest monitoring respectively. Bieling (2014) demonstrated that short written stories can be used not only as a source to detect events and species, but also in the context of Cultural Ecosystem Services. The short texts contained information about spiritual and inspirational values of landscapes, concepts related to sense of place and identity, cultural heritage and aesthetics and have many parallels with sourcing perception in LCA. Wartmann, Acheson, and Purves (2018) compared 50 texts from online hiking blogs to free-listing interviews conducted *in situ* and tags submitted to the image hosting platform Flickr. The hiking blogs contained more information about sense of place, activities and perceptual landscape aspects than free-listing interviews and Flickr tags. If we can create a reproducible workflow, capable of collecting such short texts in an automatic way, we can potentially overcome the problem of time-intensive interviews for regions where landscape descriptions are available online. Furthermore, by using text, we can retain the advantages Bieling (2014) and Wartmann et al. (2018) identified in terms of rich narrative, but collect them for large areas. Finally, if the workflow is repeatable, we can also explore how landscape descriptions vary over time, a key task in monitoring.

In this work we explore sight, sound and smell perception as well as tranquillity. References to sight prevail in both oral and written accounts of English language (San Roque et al., 2015; Winter, Perlman, & Majid, 2018), and the importance of the ways sentiments towards landscapes are expressed through language has been debated since the Romantic era introduced notions such as 'sublime' and 'picturesque' landscapes (Donaldson, Gregory, & Taylor, 2017; Herlin, 2016).

Sounds present in landscapes are perceived selectively (Fisher, 1999) with 'no direct correlation between physical measurements of loudness and perceptions of noise' (page 641, Coates, 2005). Nonetheless, a taxonomy, developed in ecoacoustics is valuable since preferences for sounds vary according to perceived emitters. For example, though natural (e.g., waterfall) and anthropogenic (e.g., jet engine) sounds may have very similar signatures, preference is expressed as a function of the nature of perceived emitters (Fisher, 1999). Three classes of sound emitter are proposed: anthrophony (sounds produced by people), biophony (sounds of animals), and geophony (non-biological natural sounds) (Krause, 2008).

To these we add perceived tranquillity, which has been shown to be a combination of sight and sound (Carles, Barrio, & De Lucio, 1999; MacFarlane, Haggett, Fuller, Dunsford, & Carlisle, 2004; Pheasant, Horoshenkov, Watts, & Barrett, 2008). One common way to classify tranquillity uses a continuous scale from least to most tranquil landscapes (e.g., Hewlett, Harding, Munro, Terradillos, & Wilkinson, 2017). However, to capture ways tranquillity is written about, we developed a taxonomy (Chesnokova et al., 2019), with four classes: combination of sight and sound, contrasting sounds, no-movement and the class of silence and tranquil sounds. As for sounds, smells are often described through emitters, e.g., 'smell of birch' (Granö, 1997; Majid & Burenhult, 2014; Quercia & Schifanella, 2015) and can be similarly classified into anthropogenic sources, and those emitted by plants or animals.

Since explicitly collected short stories and short texts available online show high potential for eliciting public perception of landscapes, our aim is to demonstrate that large volumes of written texts, retrieved from the internet, are an effective source of information about public perception towards landscapes, specifically in the context of LCA. To approach this aim, we set out to investigate the following research questions:

RQ1: How can we build a spatial referenced corpus (collection of text documents) of first-person landscape perception?

RQ2: What sorts of perception do we find in our corpus, and from whom?

RQ3: How can these results be applied for LCA?

2. Methods

To illustrate our approach, we focus on a specific case study region, the English Lake District (2.1). Using this region as an example, we describe a general workflow to collect a corpus of documents from the web, containing first-person landscape perception in the Lake District (2.2). We then demonstrate how this corpus can be analysed, extracting and classifying descriptions of sights, sounds and smells experienced by the writers of this corpus (2.3). Finally, we associate descriptions with the region as a whole, LCA areas for the Lake District (Watkins, 2008) and points associated with individual landscape elements (2.4) (Fig. 1).

2.1. Case study region

To test the potential of written textual sources we selected an area of more than 2000 km^2 in the North-West of England – the Lake District National Park (Fig. 2) – established in 1951, which became a UNESCO World Heritage Site in 2017 (Nomination, 2017). This region is not only important because of its status as a National Park and World Heritage Site, but also because of its prominence in writing about landscape and nature in English. Multiple authors (e.g., Samuel Taylor Coleridge, Dorothy and William Wordsworth) celebrated the Lakes as a place of walking and nature appreciation in the Romantic Period at the start of the 19th century (Donaldson et al., 2017). This tradition of writing has continued to the current day and now also reflects the wide range of outdoor activities undertaken there. The area is characterised by rugged topography including England's highest mountain Scafell Pike (978 m) and its deepest and longest lakes (Wastwater (74 m) & Windermere (18 km)). In the 18th–19th century the Lake District became a centre of different types of industry, including quarrying of slate, limestone, and granite (Watkins, 2008).

2.2. Creating a corpus of first-person landscape perception in the Lake District

The internet as a whole was estimated at the time of writing of this paper to contain 5.86 billion documents (de Kunder, 2019; van den Bosch, Bogers, & de Kunder, 2016). This enormous volume of natural language clearly has great potential for analysis in multiple fields. However, before analysing landscape perception, we first need to identify thematically *and* spatially relevant texts: texts containing references to first-person landscape perception *in* the Lake District. Before building a corpus we identify three key requirements. The first of these is precision – the proportion of relevant descriptions should be as high as possible. The second is recall – as many relevant descriptions as possible should be returned. The third requirement, of particular importance if we are collecting individual experiences, is that we minimise the number of duplicate, or near duplicate documents. To maximise recall, we first used a set of search terms to programmatically retrieve candidate descriptions from search engine (2.2.1). We then increased precision on this initial document set by using machine learning to classify thematically relevant texts containing first-person landscape perception (2.2.2). Spatial precision was increased by a use of a spatial filter (2.2.3) before similar documents were removed (2.2.4).

2.2.1. Initial corpus

Our initial set of candidate documents was retrieved by a Python

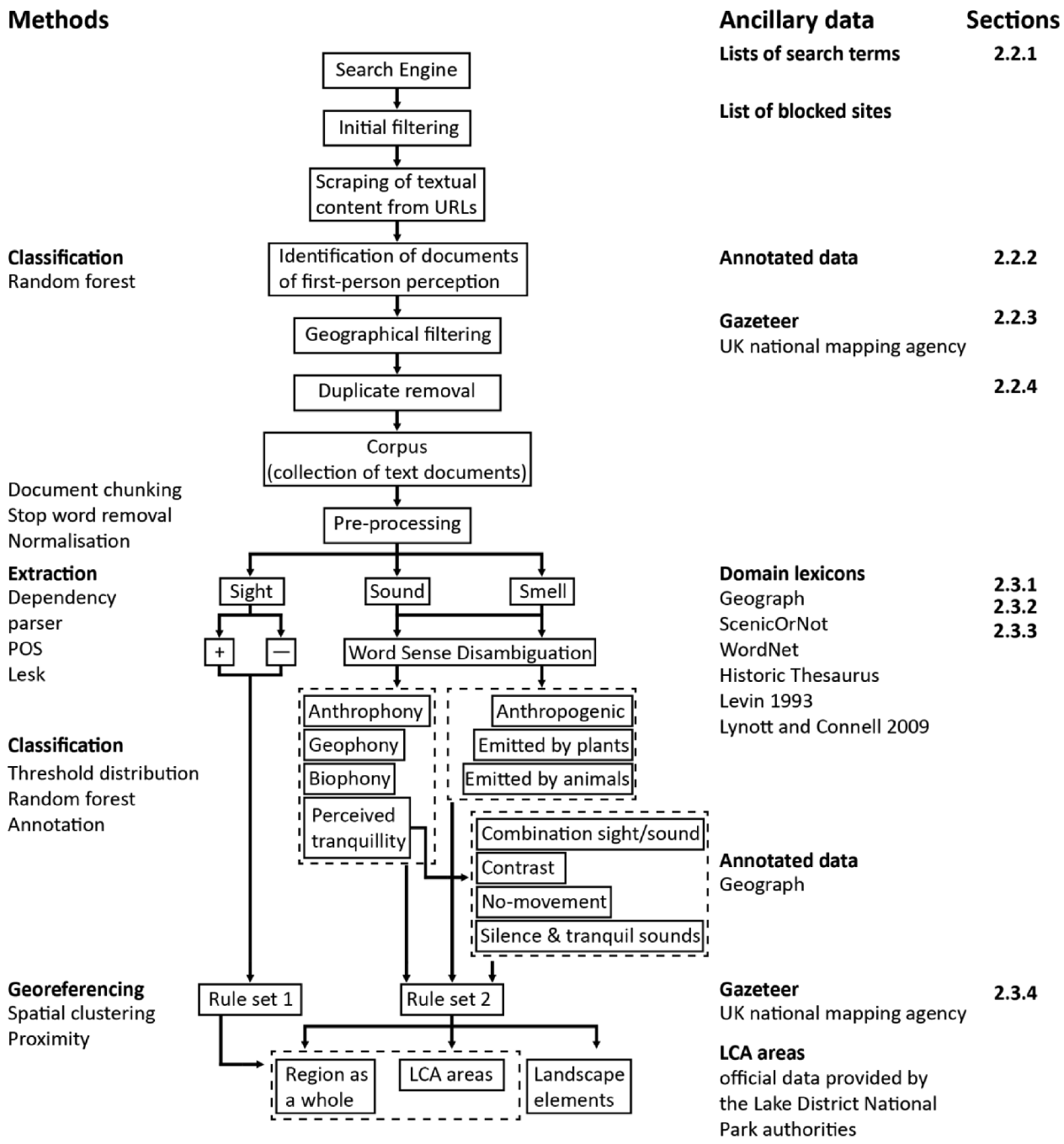


Fig. 1. Workflow including corpus creation and extraction, classification and georeferencing of first person descriptions of sights, sounds and smells.

program from the Bing search engine, using an application programming interface (API) to submit multiple queries. All queries were made with the market set to 'en-GB', specifying both preferred language and region of interest (Bing Web Search, 2019).

Each query consisted of a set of search terms likely to retrieve relevant documents (Joho & Sanderson, 2000). Initial experiments showed including "I" in the search terms increased the proportion of documents containing first-person perception. To retrieve documents relevant to the Lake District, and its landscape, we also used place names as search terms. The choice of names is central to the corpus which emerges (c.f. Davies, 2013; Wartmann et al., 2018), and we sought to address two of the categories suggested by Relph (1976) – 'behavioural insiders' and 'empathetic insiders'. 'Behavioural insiders' visit landscapes deliberately and visual patterns play a primary role. 'Empathetic insiders' do not just look at landscapes, but appreciate their identity through 'deliberate effort of perception' and understanding of 'place as rich in meaning' (page 54, Relph, 1976). To find descriptions written by 'behavioural insiders', we used a list of the names of 150 major outdoor attractions listed by TripAdvisor in the Lake District (c.f. Richards & Tunçer, 2018). These include architectural objects (e.g., castles and churches), historical landmarks, parks and gardens, viewpoints, waterfalls and houses of writers (see Appendix 1). To find descriptions of 'empathetic insiders' we used Wainwrights' list of Lake District summits, a particularly popular list for 'hill-bagging' (see Appendix 1). We assumed that those visiting such summits might more closely match the notion of 'empathetic insiders', since they may experience the landscape more intimately and more often, making many return trips to the region to collect all of the summits on the list. These lists of names can be substituted or expanded with other place names, depending on the nature of the case study region (e.g., using street

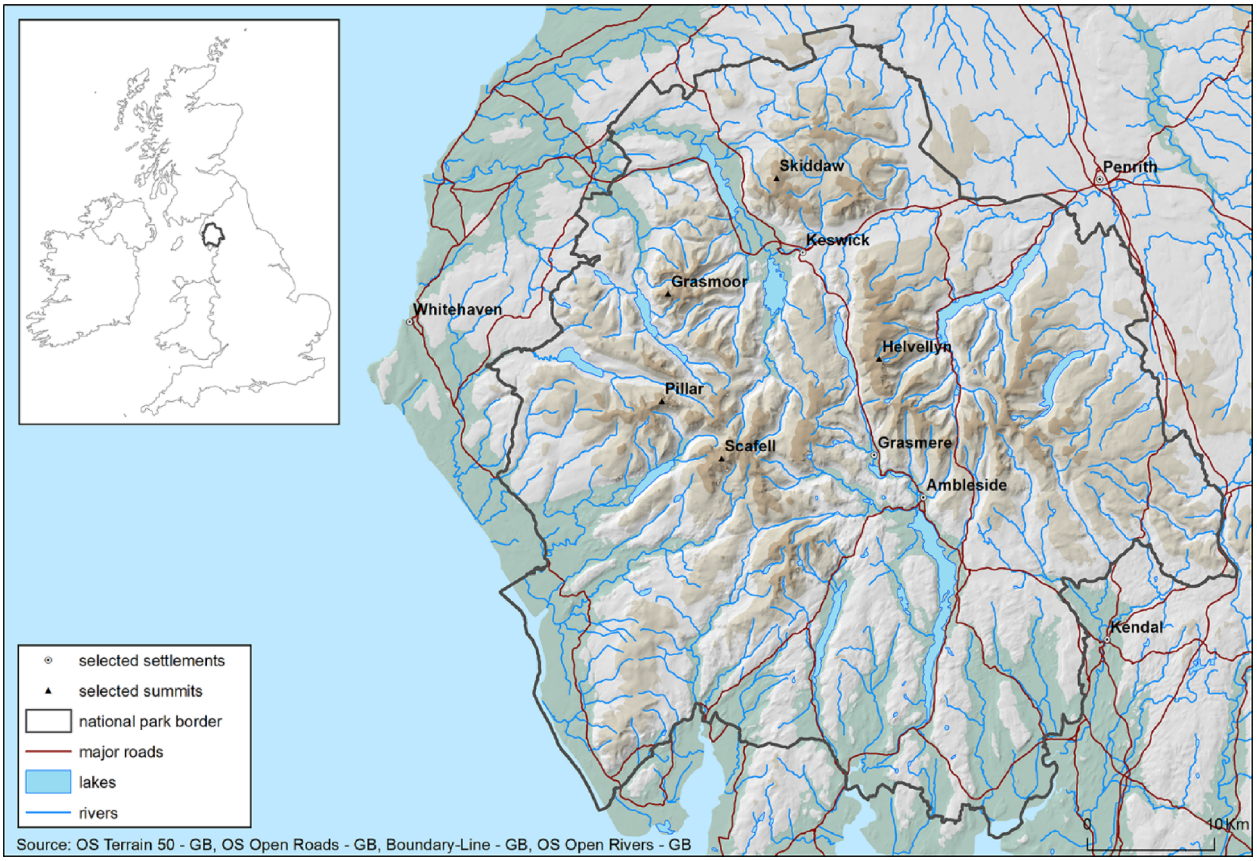


Fig. 2. The Lake District National Park and its topography.

names in an urban landscape). Place names are subject to both referent class ambiguity (e.g., Sail refers to a peak and is a common word in English) and reference ambiguity (e.g., Blencathra is also known as Saddleback) (Jones, Purves, Clough, & Joho, 2008). We dealt with referent class ambiguity by expanding our queries, adding 'Wainwright' to all searches for summits (c.f. Overell & Rüger, 2008). Reference ambiguity was dealt with by adding all known names for a given location to the search terms.

Each query returned a list of web addresses (known as URLs). Before analysing the content associated with URLs, we used a manually compiled list to programmatically remove those unlikely to contain first-person descriptions of landscape. These included words related to local government, accommodation, and Wikipedia pages, since these often contain local place names but not descriptions of individual experiences (e.g., 'gov.uk' as in https://www.lakedistrict.gov.uk/) (see Appendix). We also removed all duplicate URLs. From the remaining URLs we scraped visible textual content of all webpage elements using the Python library 'scrapy', excluding headers, footers, sidebars, and comments (Fig. 3) in accordance with the web-scraping policies of individual web sites (Greenaway, 2017; Lawson, 2015).

2.2.2. Classifying thematically relevant documents

Our search terms were designed to return documents likely to include first-person landscape perception, but a second classification step was necessary to remove false positives and increase precision. To do so, we applied a random forest, a supervised machine learning classifier well suited to textual features, using Python library 'scikit-learn' (Criminisi, Shotton, & Konukoglu, 2011; Pedregosa et al., 2011). We manually annotated training data in a preliminary study (Chesnokova and Purves, 2018a) and trained the classifier using three groups of features: the 250 most frequent words, presence of selected personal pronouns and the 50 adjectives and nouns most frequent in relevant/not relevant descriptions respectively. We split 641 annotated texts (see annotation rules in Appendix Table 1) into a 50% training and 50% test sets, and achieved precision of 0.84.

2.2.3. Filtering spatially irrelevant texts

Although we retrieved URLs using place names, these were not necessarily found in the scraped text we extracted for analysis (c.f. Fig. 3 where Rome occurs in the sidebar but not in the scraped text). We therefore performed a simple toponym recognition step using the complete list of place names used as search terms (c.f. 2.2.1) and a place name gazetteer from the UK national mapping agency for the Lake District. To account for small differences in spelling we used Levenshtein distance as implemented in Python library 'Fuzzy String Matching' (Arias, 2019), a string metric which measures how many characters need to be inserted, deleted or substituted to move from one string to another (e.g., the Levenshtein distance between cat and cars is 2 since we substitute r for t and insert s). We also used simple heuristics to match potentially compoundable nouns (e.g., Derwentwater/Derwent Water/Derwent water).

2.2.4. Eliminate similar documents

Finding duplicate descriptions contributed via different URLs is an important step, as we do not want to emphasise landscape characteristics found in multiple texts with the same source. We filtered out all descriptions with an overall string similarity of more than 80% (Python library 'Fuzzy String Matching') to create our final corpus ((Zachara & Palka, 2016; Arias, 2019; Gonzalez and Rodrigues, 2017)).

2.3. Extracting and classifying descriptions of sights, sounds and smells from our corpus

Having created a corpus of first-person landscape descriptions, we

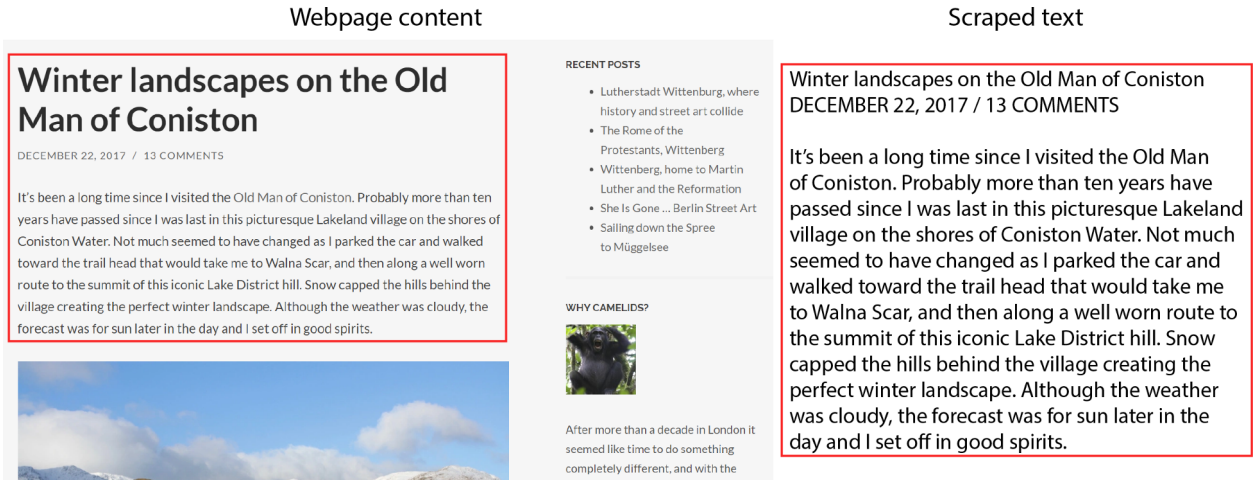


Fig. 3. Original webpage content on the left and the scraped textual content without sidebars on the right (https://notesfromcamelidcountry.net/category/coniston/, Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported Licence).

analysed experiences of sights, sounds and smells in the Lake District.

For all senses, we first performed a range of natural language pre-processing steps using the Python library 'spacy' (Srinivasa-Desikan, 2018). These included detecting paragraphs, sentences and words in documents, part of speech tagging (e.g., identifying adjectives, verbs and nouns), removal of stop words (e.g., 'a', 'the'), normalisation of words to lower case and extraction of lemmas (dictionary roots of words with the same semantic meaning) (Manning & Schutze, 1999).

After preprocessing, we automatically extracted sub-corpora containing references to sights, sounds and smells respectively. For extraction we used lexicons (lists of domain relevant terms) and pattern matching combining for example lemmas and parts of speech. For smells and sounds, we also performed word sense disambiguation, removing irrelevant uses of words (e.g., sound can refer to a body of open water).

We then further classified the extracted descriptions, in terms of scenic or unattractive elements of the visual landscape, classes of tranquillity as identified through references to sounds, and emitters for sounds and smells using a combination of machine learning and manual annotation.

2.3.1. *Extracting and classifying sights in the landscape*

References to visual perception are common in language and use a wider range of words in English than other senses (San Roque et al., 2015; Winter et al., 2018). Sentiment is often conveyed through phrases combining adjectives with nouns (e.g., compare *overcrowded summit* with *beautiful lake*), and we used this observation to guide our analysis (Liu, 2012). We were interested in collecting particularly negatively or positively connoted descriptions associated with visual perception from our corpus. This task is associated with sentiment analysis in natural language processing where lexicons are used to identify words or phrases found in, for example, positive or negative reviews (Kaji & Kitsuregawa, 2007; Lu, Castellanos, Dayal, & Zhai, 2011). Such a lexicon does not exist for landscape. To create one, we needed a collection of ratings related to landscape and texts associated with those ratings. The ScenicOrNot project (http://scenicornot.datasciencelab.co.uk/) collected more than 220,000 ratings of "scenicness" (with values between 1 and 10) for images from the Geograph project (http://www.geograph.org.uk/), a collection of representative pictures and descriptions for the whole of the UK. ScenicOrNot ratings are available under an Open Database Licence and the Geograph dataset under a Creative Commons Licence.

To build our lexicon we relied on three observations. First, since we have ratings for individual pictures and their descriptions, we can associate phrases with scenic or unattractive landscapes. Second, the Lake District is valued for its scenicness, and thus we expect unattractive descriptions to be rarer than in the UK as whole. Third, since we also know the overall distribution of scenicness ratings, we can identify phrases which are used particularly often to refer to unattractive or scenic landscapes.

Based on these observations we collected descriptions associated with scenic (mean scenicness + 2 standard deviations) and unattractive descriptions (mean scenicness − 1 standard deviation) (Fig. 4). Doing so resulted in 4847 scenic and 26,029 unattractive descriptions for the UK as a whole.

To create our lexicon of phrases associated with unattractive and scenic landscapes, we then extracted adjectival modifiers using a dependency parser (e.g., from the phrase, 'stunning panoramic views', we extracted two pairs: 'stunning views' and 'panoramic views') (Honnibal & Johnson, 2015), and tested these for significance compared to all descriptions (Chi-square test, df = 1, $p < 0.005$). We only retained phrases which were associated with particularly high or low ratings of scenicness and not simply common overall. The resulting lexicon contained 184 scenic phrases and 214 associated with unattractive descriptions (see Appendix).

2.3.2. *Extracting and classifying sounds in the landscape*

To extract descriptions related to sounds in the landscape, we also used a lexicon. We took a top-down, knowledge-based approach, and

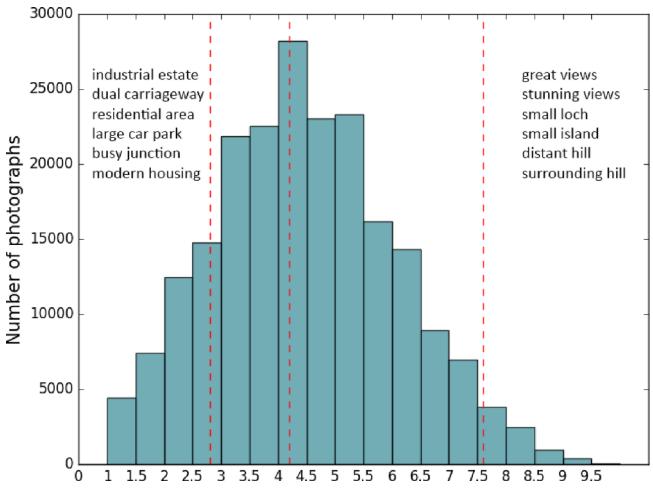


Fig. 4. Distribution of scenicness values for all pictures and descriptions, thresholds for scenic and unattractive descriptions and examples of adjectival modifier pairs extracted.

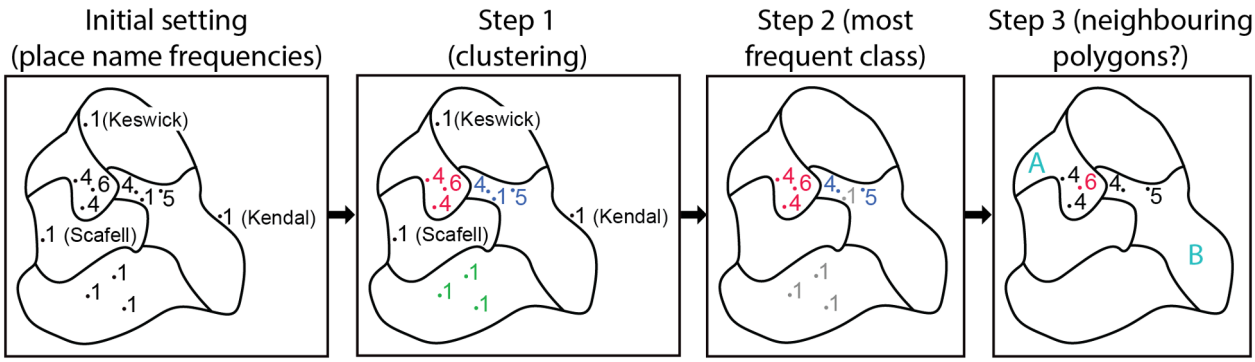


Fig. 5. The workflow to assign documents to LCA areas. Different colours in Step 1 correspond to different clusters, blue and red locations in Step 2 belong to the most frequent class, and red location with frequency 6 in Step 3 is the most frequent place name. Areas A and B are LCA areas associated with the description. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

built a lexicon consisting of English verbs of sound emission, sound existence and sounds made by animals (Levin, 1993), and synonyms for all these verbs according to WordNet (Fellbaum, 1998). We added a list of adjectives related to sounds. Since our previous work had shown that silence is an important sound related quality in the landscape (Chesnokova et al., 2019), we added a list of terms referring not only to perceived absence of sound, but to tranquillity as a holistic combination of sight and sound experiences. We took terms from the Historical Thesaurus of English (https://ht.ac.uk/) in the categories "inaudibility," "faintness/weakness" and "quietness/ tranquillity" to create the final lexicon consisting of 262 sound and tranquillity related words (see Appendix).

We then extracted candidate descriptions containing sound-related first-person perception by matching lemmas of lexicon terms to our corpus. This generated many false positives, since words describing sound experiences are highly polysemous (e.g., echo can be used literally with respect to sound or figuratively to describe styles). We disambiguated such cases by, first, controlling for the correct sense of verbs and nouns using WordNet categories as implemented in the Lesk algorithm (Manning & Schutze, 1999). Second, we developed rules using part of speech tagging (e.g., taking into account 'still' only if it is labelled as adjective as in 'still waters' and not as adverb as in 'still hungry'). Finally, based on examination of false positives, we added additional rules to filter common ambiguities (e.g., removing 'screaming calves', usually referring to muscular pain).

We used an existing taxonomy of sound emitters, classifying sounds as biophony, anthrophony and geophony (Chesnokova and Purves, 2018b), adding an additional class often found in first-person landscape descriptions, absence of sound, to account for descriptions conveying a general sense of peace in terms of sounds and sights (c.f. Pheasant & Watts, 2015). We manually annotated 8784 descriptions in the Geograph collection according to this taxonomy with a Cohen's Kappa inter-annotator agreement of 0.88 (Landis & Koch, 1977). After exploring these texts in detail, we redefined absence of sound to include perceived tranquillity (see Appendix, Chesnokova and Purves, 2018b).

Using these annotated data, we implemented a random forest classifier using Python library 'scikit-learn' (Criminisi et al., 2011; Pedregosa et al., 2011), using as features the 500 most common words, a list of British birds and mammals and a list of natural elements and related qualities to classify extracted and disambiguated sound descriptions as either biophony, anthrophony, geophony or tranquillity (see Appendix), and achieved precision of 0.81.

Since tranquillity can be usefully and reliably classified by human annotators, we sub-classified each description referring to tranquillity to one of the following 4 categories: contrasting sounds; combination of sight and sound perception; no-movement and total silence and tranquil sounds (Chesnokova et al., 2019) (see annotation rules in Appendix Table 2).

2.3.3. Extracting and classifying smells in the landscape

The final sense extracted from our corpus was smells. As for sound, we created a lexicon based on verbs of smell emission (Levin, 1993) extended by WordNet lists of olfactory categories and adjectives with dominant modality "olfactory" (Lynott & Connell, 2009). This lexicon contained 29 words (see Appendix). We disambiguated candidate descriptions using an analogous process to that performed for sound. Finally, since references to smells were relatively rare, we classified these manually into those emitted by plants, animals and anthropogenically.

2.4. Associating classified descriptions with space

Having extracted and classified descriptions associated with sights, sounds and smells, we were left with a subset of documents containing relevant descriptions of individual senses. For each of these descriptions, we could identify the sentence related to landscape perception and an attribute indicating associations with senses and the resulting classification. An individual description could be associated with one or more senses.

These descriptions could be analysed without any further processing, as characterising the Lake District. However, our motivation was to provide descriptions relevant to LCA, and to do so we had to explicitly link text to space. We used place names found in the texts to link sight related descriptions to LCA areas (Watkins, 2008) and sounds and smells to the point locations of place names (e.g., summits or settlements) found in their descriptions.

To assign sight-related documents to LCA areas we performed three steps having initially calculated place name frequency in each text (Fig. 5). First, we applied density-based clustering as implemented in PostGIS (Moncla, Gaio, & Mustière, 2014; PostGIS, 2019) to disambiguate and detect outliers of seen but not visited locations. Second, we created three classes of place name frequency based on Jenks natural breaks data clustering (Dara-Abrams, 2011) and retained only the most frequent class (which we assume to be more likely to be visited and thus experienced). Finally, we took the most frequent place name and performed a region-based disambiguation on the other place names found in the most frequent class. All steps related to spatial analysis were performed using Python library 'arcpy' (Esri, 2019).

For sounds and smells, we first looked for a place name in the relevant sentence, checking for referent ambiguity (does this place name occur more than once in the Lake District?). If not, then we assigned the coordinates found in a gazetteer. In cases of referent ambiguity, we disambiguated using other place names found nearby in the text using a distance-based measure (Leidner, Sinclair, & Webber, 2003).

Finally, to reduce the effects of bias induced by participation inequality (Nielsen, 2006), we retained only one description if several had the same class and location and were generated by the same user.

Having performed these steps, we have a list of perceived landscape

Table 1
Corpora statistics.

	TripAdvisor	Wainwright
Number of search terms (including different spellings)	92	214 (233)
Initial number of extracted texts	13,110	34,150
Number of relevant texts in the Lake District	961	5909
Average paragraphs per text	49	79
Average sentences per text	81	104
Average words per text	1277	1120

properties associated with LCA areas (sights) and individual landscape features (sounds and smells).

3. Results and interpretation

3.1. Thematic corpus of the Lake District

Using our customised lists, we initially retrieved 13,110 and 34,150 texts, for search terms derived from TripAdvisor locations and Wainwrights' list respectively. After the filtering stages described in Fig. 1 we were left with a total of 6870 relevant texts and a corpus consisting of almost 8 million words (Table 1). Documents varied in the nature of the information they contained ranging from descriptions of participation in a single event (e.g., a hike to a summit) through multiple descriptions of different locations by the same individual or collections of descriptions of a single location from multiple users. Thematically, contrary to our expectations, both sets of texts contained a broad mix of activities and narrative types without a clear distinction between 'behavioural insiders' and 'empathetic insiders' (Relph, 1976) and we treated texts thereafter as a single corpus of first-person perception (Table 1).

To investigate the efficacy of our lists in retrieving spatially relevant texts, we report here on our ability to link documents to LCA areas. For the 71 LCA areas in the Lake District, we could collect more than 10 texts for 54. Only a single area – Lyth Valley – has no texts. Peripheral areas, and in particular the southern part of the Lake District, not described in Wainwright's list, have fewer texts. Unsurprisingly, more texts are found for famous parts of the region, containing both the high mountains of Scafell and the popular valley landscapes around Grasmere (Fig. 6).

3.2. Characterisation of perceived landscape properties

An important challenge in the development of new methods to extract information is their potential utility for practical applications. The challenge of assessing whether or not we extract useful information is well-known, and we are interested in identifying *interesting patterns*. These are characterised in the Knowledge Discovery domain by unexpectedness (a user learns something new) and actionability (a user learns something upon they might act) (Silberschatz & Tuzhilin, 1996). To evaluate these, and other properties of our results, we created a set of web maps which we used in subsequent interviews with experts in the Lake District (3.3). For sight, we generated word clouds for the 34 LCA areas with two or more texts per km^2 (c.f. Fig. 6) and scaled the top 50 scenic and unattractive adjective modifier pairs using spatial term frequency/inverse document frequency (Rattenbury & Naaman, 2009) to identify locally distinctive pairs (Fig. 7). For sounds and smells we visualised individual descriptions as points with extracted sentences, paragraphs and the original URL available (Fig. 7).

In the following, we give selected examples of how these maps can be used to characterise our region of interest – the Lake District – as a whole and at the spatial scale of the LCA areas. In doing so we move from macro- to micro-analysis by looking at emerging patterns of automatic analysis in the former and zooming in to interpret individual descriptions to better understand these patterns in the latter (Jockers, 2013).

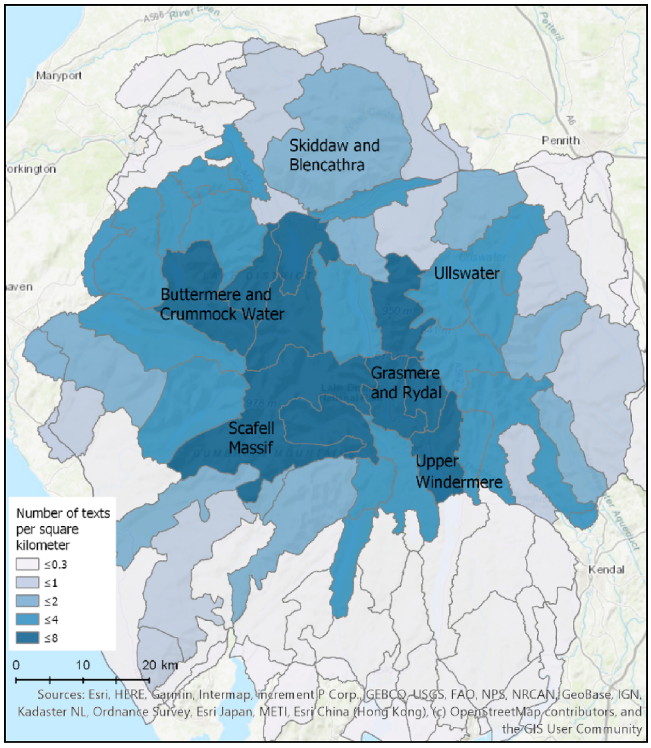


Fig. 6. Document density for LCA areas.

3.2.1. Characterisation of perceived landscape properties on the scale of the Lake District

Using our domain specific scenicness lexicon, we extracted a total of 28,179 descriptions referring to scenic landscapes and only 266 descriptions referring to unattractive locations. These values are in themselves not surprising, since, first, our lexicon is based on the whole of the UK, and, second, the Lake District is characterised by its outstanding scenic qualities. Table 2 lists the ten most commonly occurring pairs found.

Many of the terms associated with scenic locations relate to generic visual properties such as great, good and stunning view(s). More experiential perception related to locomotion in the landscape is also common, as for example, steep ascent, steep descent and good path. These can be related to Wainwright, who considered 'bodily experience' an important component of landscape perception (Palmer & Brady, 2007). Zooming in to individual descriptions demonstrates the importance of this dimension for the writers in our corpus: "When we finally met up with the Stake Pass and could head down hill on a *good* solid and visible *path*, it was truly time to celebrate." (https://ramblingman.org.uk/walks/wainwrights/southernfells/esk_pike). Although unattractive elements are uncommon, they are also revealing, relating to negative sentiments towards transport (e.g., large (car) park, parked cars and dual carriageway): "Tarn Howes is a special place, albeit too close to a *large* car *park* and not seen at its best because of the poor light and lingering mist." (https://lonewalker.net/walkinfo.php?walk=412). Other references to unattractive elements refer to previous industry (e.g., old works, old machinery): "Dotted all around are spoil heaps, rusting iron cables lie along the path, bits of *old machinery* lay abandoned on the mountainside, and a metal tower from an aerial tramway lays toppled on its side." (https://notesfromcamelidcountry.net/category/coniston/). Such abandoned mines and quarries are common in the Lake District, however, they are often ignored by the writers of our corpora, following the tradition of William Wordsworth,

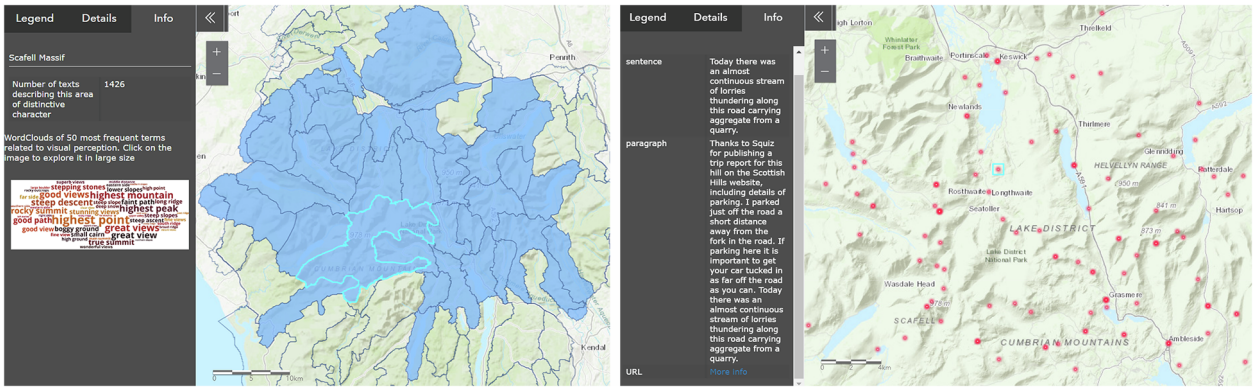


Fig. 7. Interface of the web maps demonstrating the results of our approach (tinyurl.com/LakeDistrictPerception). Left: Word cloud displayed for the selected LCA area "Scafell Massif". Right: Map of sound descriptions classified as anthrophony and an example description.

Table 2
Ten most frequent combinations from scenic and unattractive lexicons found in our corpus.

	Scenic pairs	Count	Unattractive pairs	Count
1	great views	1012	large (car) park	39
2	highest point	881	parked cars	26
3	good views	707	dual carriageway	18
4	steep descent	528	old works	12
5	steep ascent	370	old machinery	10
6	good path	353	adjacent park	6
7	good view	315	local shops	6
8	stunning views	314	static caravans	6
9	great view	306	much traffic	6
10	lower slopes	296	second bridge	5

Table 3
Summary of extracted descriptions of sound experiences per class.

Type of sound experience	Count
Combination of sight and sound perception	485
Contrasting sounds	275
No-movement	66
Tranquil sounds and total silence	60
Total perceived tranquillity	**886**
Anthrophony	174
Biophony	142
Geophony	278
Total assigned to emitter	**594**

who deliberately overlooked not only the appearance of – at this time still functioning – mines, but also the sound of the excavations (Taylor, 2018). Indeed, the majority of the sound related descriptions (886 out of total 1480 descriptions) refer to perceived tranquillity and the general absence of sound (Table 3).

Most common amongst such descriptions are those referring to a combination of sight and sound perception as in "The walk back to Wasdale Head is largely along the road but it is fairly quiet and offers ample opportunities to leave it to admire the views." (http://allthegearbutnoidea.blogspot.com/2018/06/a-circuit-of-wast-water.html). Contrasting sounds, which typically characterise a location favourably by comparison to, or despite, a nearby place's properties are also prominent as in this description of Broadcrag Tarn: "When the summit of Scafell Pike is crowded with excited and chattering groups of walkers, it is a place to visit for here very few people tread and apart from the occasional curious Herdwick sheep you are unlikely to meet a soul." (https://herdwickcountry.wordpress.com/2014/07/08/broadcrag-tarn/). These results reveal an important property of our work: having classified descriptions we can also perform micro-analysis to start to understand how the Lake District is experienced. In the class no-movement, the influence of the Romantic poets continues to persist, with its emphasis on the mirror-like reflection of lakes and tarns as in "We made a slight detour which brought us closer to Hayeswater, which was now free from mist and mirror like in its stillness." (https://these8boots.wordpress.com/category/patterdale/). These descriptions can be seen as not properties of individual locations, but rather an important part of the eponymous identity of the Lake District related to the water of its lakes and tarns. The influence of Wainwright's writing is also reflected, his favourite places Haystacks and Innominate Tarn appear in several descriptions of tranquillity, e.g., "In absolute silence we were treated to the reflection of the sky and the distant heights of Pillar in the totally flat waters of the tarn" (http://adventuresforthecommonman.blogspot.com/2018/04/sometimes-its-about-journey.html). As well as analysing tranquillity, we also produced maps of specific sound emitters: for anthrophony, rumbling traffic/traffic noise and for biophony, roaring/bellowing stags (Fig. 8). These demonstrate the relationship between different experiences and the Lake District as a whole, as well as specific locations. Thus, traffic noise is experienced and written about not only in the valley, but also on summits close to the major transportation axis of the A591, as for example on Helm Crag, "As you pass around to the other side of the fell the road noise from the A591 is more audible." (http://mydadsboots.blogspot.com/2015/10/the-lake-district-helm-crag-and-gibson.html). A distinct soundscape emerges from our texts as a cluster of locations south of Ullswater characterised by roaring stags: "This is the oldest red stag area in England and the "Rut" had began as roaring stags could be heard all around." (http://frasermackay.blogspot.com/).

Though smell perception is rare in English in comparison with sight and sound perception (Majid & Burenhult, 2014), smellscapes are still important in defining the unique character of places (Dann & Jacobsen, 2003). From our corpus, we identified 78 smell descriptions, 48 of which describe smells emitted by plants, 21 by anthropogenic sources (e.g., food, smoke) and 7 by animals. Anthropogenic descriptions capture popular tourist locations such as Grasmere's gingerbread shop, but also the common smell of burning brakes on the steep Hardknott pass: "Cars on 3 wheels coming around the steep hairpins, stinking of burning clutches and brake discs." (https://babtestingground.wordpress.com/2012/10/). We find dotted around landscape references to the scent of blossoming heather, sweet gorse and the pungent smell of dead sheep, often near steep Lakeland cliffs. As for sound perception, smell experiences reflect both temporally dynamic (e.g., blossoming flowers) and spatially variable (e.g., the final resting places of the unfortunate dead sheep) processes, but also encode information about the affordances of particular landscape elements (such as the heather covered slopes of Blencathra).

3.2.2. *Relating perceived landscape properties to LCA*

LCA has, at its core, the production of narrative descriptions related

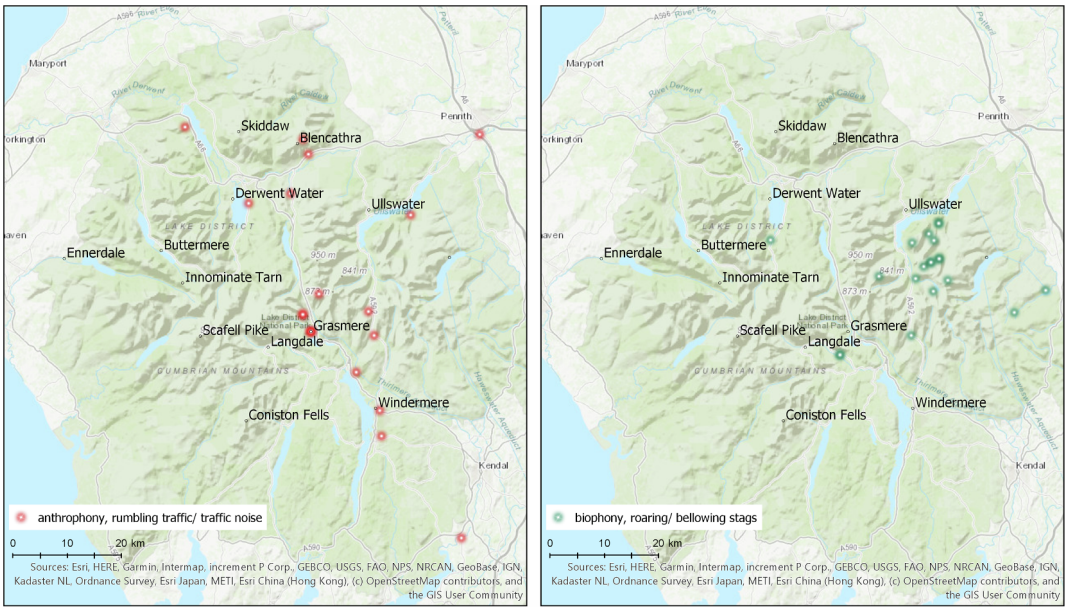


Fig. 8. Selected examples of sound experiences spatial distribution. Left: anthrophony, rumbling traffic/traffic noise, 19 descriptions. Right: biophony, roaring/ bellowing stags, 20 descriptions.

to defined areas. Our extracted texts provide insights into perceived elements of sight, sounds and smells, making it possible to characterise landscapes at the level of the LCA areas, identify areas having similar characteristics in the region as a whole and compare them to their neighbours.

Our first macro-analysis exploration reveals that only 13 out of 34 LCA areas contain pairs from the unattractive lexicon. Two neighbouring areas: 'Helvellyn Range' and 'Brother's Water and Hartsop' are characterised by 'large (car) park', however, here the sentiments are not negative as in the example of Tarn Hows (c.f. 3.2.1), but reveal a more complex interplay between unattractive sight perception and overall positive sentiment: "Seventeenth century Hartsop village […] a lovely little place playing host to a rather *large* car *park*, an ideal starting point for the ascent of High Street." (http://www.one-foot.com/Over%20High%20Street%20return%20through%20pasture%20Beck%20Bottom%202012.html) and "The Helvellyn range is well served by a *large* car *park* in Glenridding" (http://allthegearbutnoidea.blogspot.com/2013/07/helvellyn-via-striding-edge.html).

For sound perception we calculated the proportion of texts describing sound experiences (in general and per class) to the total number of texts describing the corresponding area (Table 4). Looking at both sight and sound demonstrates that, for example, 'Upper Windermere' area has not only unattractive pairs, capturing its urban characteristics such as busy carriageway, parked cars and modern estate, but it also has the highest proportion of anthrophony related to traffic and noise referring to Royal Air Force training activities: "RAF trainer buzzing Kirkstone Pass" (http://www.loweswatercam.co.uk/130219_To_Sweden_with_Two_Pikes.htm). 'Claife Heights and Latterbarrow' area also contains several unattractive pairs, but sound is not characterised by anthrophony, but rather by its absence through contrasting sounds (13 of 23 tranquillity descriptions): "I have been through Kirkstone Quarries before and it is usually a hive of activity, but today it is uncharacteristically quiet." (http://www.flamingonion.co.uk/langdale_walk/). This text was written in 2012. In 2014 another author writes that the quarries have closed and "Now the slate cutting rooms and showrooms stand quiet and empty." (http://tandjinlakes2014.blogspot.com/), showing the potential of our approach to document change.

Two neighbouring areas 'Skiddaw and Blencathra' and 'Keswick and Derwent Water' are dominated by tranquillity related sound experiences (Table 4). However, in the 'Keswick and Derwent Water' area total silence is a scarce resource (only 1 mention) in comparison to 12 reports for 'Skiddaw and Blencathra'. 'Keswick and Derwent Water' is additionally characterised by high proportions of all other types of sounds: anthrophony, geophony and biophony. Zooming in we see that in contrast to 'Upper Windermere', with its traffic noise, anthrophony for 'Keswick and Derwent Water' is characterised by chugging boats and noises of other visitors. Two important tourist locations Lodore Falls and Ashness Bridge are within the borders of this area giving "sweet sound" (https://insearchofbritain.wordpress.com/tag/robert-southey/) and "wonderful sound" (https://upnoutside.wordpress.com/tag/ashness-bridge/), respectively. We also find references to pleasant sounds associated with wildlife (biophony), above all in the form of birds.

Despite the relative rarity of descriptions of smell perception, these can suggest important properties of the area as a whole, as for 'Skiddaw and Blencathra' where 4 of 9 descriptions mention the scent of "heather in full bloom" (https://www.wainwrightwalking.co.uk/ullock-pike-to-dodd/). Contrasting properties of the LCA areas to their neighbours confirms again the anthropogenic nature of 'Upper Windermere' with a single description referring to the anthropogenic smell of fish and chips, while the neighbouring area of 'Grasmere and Derwent Water' has seven descriptions, with four relating to a diverse range of plants including juniper, magnolia and hyacinths.

3.3. Expert group discussion

To evaluate the potential of our methods and its results, we visited the Lake District National Park authority for an expert discussion with, on the one hand, the authority itself, and on the other an important local lobby group (the Friends of the Lake District). We prepared a short presentation to introduce our approach, the web maps described above (tinyurl.com/LakeDistrictPerception) and a structured set of questions to discuss the utility of our approach based around a SWOT analysis (Strengths, Weaknesses, Opportunities and Threats). This study was not designed to provide a comprehensive evaluation, but rather feedback as to potential in a practical setting. Three participants took part in the expert discussion, and we summarise their feedback in Table 5.

A number of important points emerged from these discussions. The expert group explicitly saw the potential utility of such narrative descriptions for LCA, and also noted the value of a repeatable method for

Table 4
Five highest numbers of descriptions per type of sound emitter normalised by the number of texts describing corresponding area (in the brackets the absolute frequency is given).

	All sound experiences	Tranquillity and absence of sound	anthrophony	geophony	biophony
1	Skiddaw and Blencathra (117)	Skiddaw and Blencathra (82)	Upper Windermere (8)	Ennerdale (17)	Shap and Birkbeck fells (6)
2	Keswick and Derwent Water (74)	Claife Heights and Latterbarrow (23)	Keswick and Derwent Water (8)	Coniston Fells (13)	Martindale (8)
3	Coniston Fells (53)	Coniston Fells (33)	Borrowdale (11)	Keswick and Derwent Water (16)	Brother's Water and Hartsop (12)
4	Claife Heights and Latterbarrow (34)	Keswick and Derwent Water (42)	Thirlmere (6)	Borrowdale (22)	Keswick and Derwent Water (8)
5	Ennerdale (50)	Buttermere and Crummock Water (54)	Ullswater (10)	Grasmere and Rydal (18)	Kentmere Fells (9)

gathering such data. The importance of including detailed temporal information for monitoring was a key weakness, and a challenge that we discuss more below. In general, when working with subjective information, the participants pointed out the importance of context with respect to who had provided information, something also suggested as a weakness of current (expert-dominated) approaches to LCA (Butler, 2016). However, our experts also identified actionable patterns which we had both expected (relating landscape qualities more directly to perception) and which surprised us (identifying important indicator species not perceived by visitors). The potential utility of our tools in management, reflects broader initiatives in better understanding visitor behaviour in protected areas through such novel sources (Toivonen et al., 2019).

4. Discussion

In the introduction we set out aims, which can be summarised as methodological (how can we build a spatial referenced corpus of first-person landscape perception?), thematic (what sorts of perception do we find in our corpus, and from whom?) and potential (how can these results be applied for LCA?). We now discuss each of these questions in turn, pointing out not only strengths and weaknesses of our approach, but also more general implications for studies of landscape.

Edwards (2018) suggested that creative writing 'on, and better still *in* a landscape' (page 666) makes people focus on the senses and feelings these landscapes evoke, and that by sharing these personal stories people demonstrate their care for a particular landscape. She argues for inclusion of creative writing practices in the process of LCA. In this work we join her call, developing a customisable and repeatable workflow, to collect descriptions of first-person landscape perception, which we see as creative writing contributions published online. We used these texts to look at the ways LCA can capture multiple voices and become less expert-dominated.

We aimed for high precision (i.e., the descriptions we identify are likely to truly contain first-person perception) at the cost of missing other, potentially relevant descriptions. By using existing search engine APIs and lists of search terms we can rapidly build, filter and extract descriptions of specific locations. The resulting corpus contains almost 7000 individual texts and around 8 million words. By way of comparison, Bieling (2014) built a rich corpus of 42 short stories using more traditional participative methods. To demonstrate transferability we repeated the first two steps of our overall workflow (Fig. 1) for another national park in England – the Broads – since its geographical characteristics deviate strongly from the ones of the Lake District. The Broads is a flat region in the East of England with an exceptionally developed navigable network of rivers and lakes used for sailing. Using entries of the UK national mapping agency gazetteer spatially located within the borders of the Broads from which we removed entries referencing to farms and houses (total of 199 unique entries) and 26 entries from TripAdvisor we extracted 40,402 unique URLs excluding blocked sites (step 2, Fig. 1). Thus, we are confident that our approach is applicable to different landscapes.

Our corpus is well distributed across the whole Lake District and shows a strong relationship to the initial distribution of search terms, suggesting that customising lists would enable us to return more documents. Customisation brings us to a first important implication for landscape research more generally. Methods seeking to classify text remain dependent on lexicons and training on domain specific texts is essential for the development of new methods. As applications of text analysis in landscape research grow, there is a need to develop customised resources and methods for landscape research. We emphasise that our methods have been developed and trained on data in the English language, and we do not expect all results to be culturally invariant (Mark & Turk, 2017). Our texts are produced by a self-selecting group of individuals, who reflect neither all experiences nor opinions about the Lake District. Language itself is biased towards positivity

Table 5
Summary of key questions posed and feedback from expert discussion.

Question	Feedback
Strengths:	
What do you see as particular strengths of this approach and the results presented?	The methods are *repeatable and automatic*, making *comparison*, for example, between each 5-years periods *possible*. The *value* of the approach for *Landscape Character Assessment* was *explicitly stated*. Possibility to *change search terms* makes *method robust*.
Weaknesses:	Sources may be *biased* to *positive descriptions*.
What weaknesses do you see with respect to the approach and the results presented?	*Missing temporal information* is very important for *monitoring* and estimating *differences due to seasonality*. *Peace and tranquillity* may be dependent on *(unknown) background* of writer. Descriptions are not *stratified* according to either *activity* or *experience* (e.g. looking at versus visiting the summits)
Opportunities:	Monitoring of *landscape quality* and relation to *perception of visitors*. Monitoring of *opinions* towards *access management* actions and other *planning decisions* (e.g., before/after). Identification of *topics and species* (e.g., alpine plants on Scafell), which are *not perceived*, but are *important ecologically*. Such topics could be used in *visitor education*. Incorporating *other sources* (e.g., Twitter, Facebook groups) could show *more negative and instantaneous opinions* (e.g., traffic jams during the Bank Holidays).
How could your organisation use this approach and these results, if at all?	
Threats:	
What dangers do you see in adopting these methods, and in working with these results?	*Potential misrepresentation* of *certain groups* of users (e.g., hill climbers).

(Dodds et al., 2014), and individuals are more likely to report on positive experiences in landscapes (Taylor, Czarnowski, & Flick, 1995). However, by comparing what is reported across a region, we can still make an important contribution to understanding landscape perception. Finally, since we rely on scraping of content, we note the importance of considering ethical issues in so doing (Zimmer, 2018).

Our second contribution builds on our corpus, to analyse and classify different forms of perception (sight, sound and smell) at different scales and using a range of methods. We transferred a random forest classifier, which identified different sound experiences, directly to these texts, demonstrating that our approach is robust. However, we also had to rely on manual annotation to classify rare descriptions (such as those related to smell perception). Despite the emergence of other approaches which classify potential perception based purely on the existence of emitters (e.g., Quercia & Schifanella, 2015), we focussed on experienced (and not potential) perception. Our workflow, extracted and classified 28,445 descriptions referring to sight, 1480 descriptions sound experiences and 78 describing smells. This information is valuable at a range of scales, for example, allowing us to characterise and compare the nature of tranquillity in LCA areas. Our methods rely on linking perception to locations and coordinates, where improvements are still required. For example, we can only separate visited from seen locations in a rudimentary way (c.f. Moncla, Renteria-Agualimpia, Nogueras-Iso, & Gaio, 2014) and we treat all landscape elements as point locations.

The European Landscape Convention, and our expert group, emphasise the importance of landscape as perceived by people, and in turn, we need to know something about who describes landscape. Although we expected our two lists to capture different sorts of users – 'behavioural insiders' and 'empathetic insiders' – this difference turned out to be less clear cut in practice. Furthermore, in traditional monitoring instruments demographic information such as age, gender, and occupation are considered important (Kienast, Degenhardt, Weilenmann, Wäger, & Buchecker, 2012), and for all of these we have no information. We suggest two possible directions here. First, as with the boom in research on social media, there is a need to reflect on ways of modelling who is active in the landscape, who writes about it, and how we can classify characteristic behavioural patterns (c.f. Komossa, van der Zanden, Schulp, & Verburg, 2018). Second, many of our texts contained detailed information about those producing the content. There is no reason why, with appropriate ethical approval and data protection, that such writers cannot be approached and surveyed to reveal more about those producing such data (c.f. boyd, 2007).

Our last question concerned the potential of our approach. First, we hope that the rich examples we have produced exploring the ways in which the Lake District is perceived demonstrate clearly how texts can be extracted, classified and analysed. We see potential of our approach, for example, for 'landscape biography', as here first-person historical narratives are also important (c.f. Kolen & Renes, 2015) and our extraction techniques can contribute to the plurality of collected stories, since the methods could be adapted to historical written accounts. We did not look specifically at the extraction of memories in our work, but covered the aspect of soundscapes important for identification of historical patterns in landscapes (Kolen, Renes, & Bosma, 2018). Second, we evaluated and discussed our approach with an expert group. Although this group was small, they were able to identify and suggest ways such data could be used in practice, showing that our approach has practical utility. Indeed, for monitoring, we see such texts as an interesting way of predicting potential change. For example, one description explicitly comments on the tranquillity of a hill recently resurveyed with a height of more than 2000 feet (and thus reclassified as a mountain!). An online news report comments, "Miller Moss is not the most exciting hill in the world but it should become a little busier now" (Barnard, Jackson, & Bloomer, 2018). Tracking how this landscape element is described in future writing could illustrate if this prediction comes true, but would also require us to more systematically analyse time. To explore temporal change in our corpus, we used the temporal tagger HeidelTime (Strötgen & Gertz, 2010) to extract references to dates. Fig. 9 shows that almost 50% of the texts we analysed are from 2017 and 2018, suggesting that texts describing first-person perception appear to have a rather short period of existence, pointing to the importance of archiving such material if it is to be used for research in the

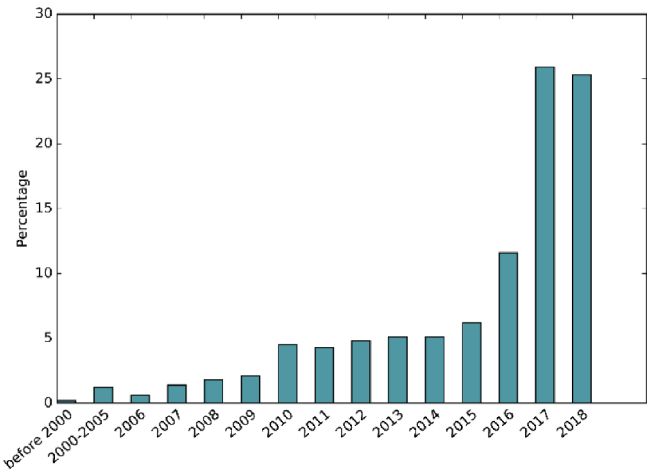


Fig. 9. Temporal distribution of collected texts.

future (Hale, Blank, & Alexander, 2017).

Of course, our approach cannot replace traditional approaches to LCA. Rather we see it as a way of exploring, across large volumes of texts, the diverse ways in which landscape is experienced, and providing impetus to other methods, which seek to better incorporate such perception in LCA and similar approaches.

5. Conclusions and further work

Our starting point was the need to integrate perception, as experienced by those visiting a landscape, in landscape characterisation and monitoring. Inspired by the narrative nature of LCA, and the importance of incorporating different viewpoints and senses, we used the internet as a source to extract and analyse texts capturing first-person perception in the Lake District. Our results demonstrate that, it is possible to build a large, diverse corpus of first-person landscape experiences, and analyse it with respect to multiple senses. More profoundly, they demonstrate that text is a rich, though as yet rarely exploited potential source of landscape information. To utilise such information there is a need to develop landscape specific methods for analysis, which are culturally and linguistically sensitive, and to rethink oversimplistic taxonomies of landscape use. In our texts we find multiple, intertwined experiences, which cannot be meaningfully disentangled into tourist and local perspectives. Text also lays bare the influences of other discourses on the ways in which landscapes are experienced, and reaffirms the importance of considering how guidebooks, modern and historic nature writing, and poetry can influence ways in which landscapes are perceived and remembered (Prior, 2017).

Analysing such information requires that we develop effective approaches to identifying both widely shared views, and also more marginalised opinions, which may capture groups not well represented in the underlying data. Equally, our rich corpora would lend itself to a wide range of other qualitative and quantitative analyses, for example, exploring sentiment with respect to landscape, or perception related to biodiversity indicators.

CRediT authorship contribution statement

Olga Koblet: Conceptualization, Methodology, Software, Validation, Visualization. **Ross S. Purves:** Conceptualization, Methodology.

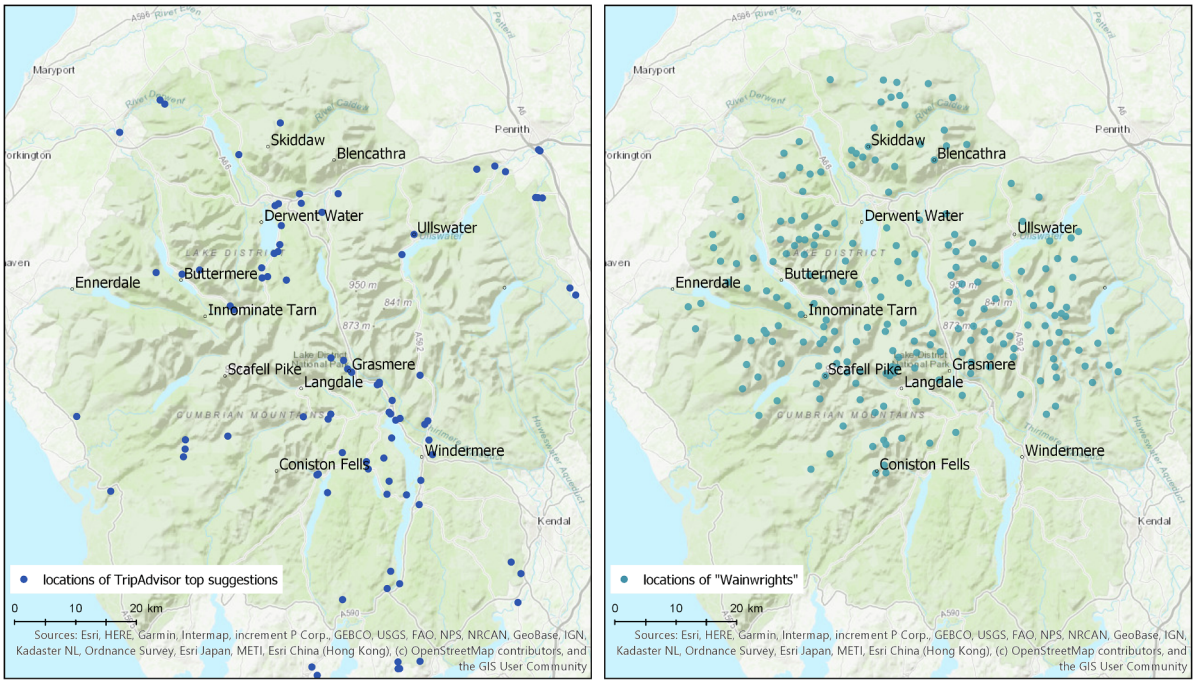
Acknowledgement

We would like to thank all the contributors to ScenicOrNot (Open Database Licence) and Geograph British Isles (Creative Commons Attribution-ShareAlike 2.5 Licence), local experts for their valuable feedback and the reviewers for their insightful comments, which improved the clarity of the article. RSP gratefully acknowledges support from the Swiss National Science Foundation Project EVA (200021E-166788) and the University Research Priority Programme on Language and Space.

Appendix

Terms used for the initial filtering of the returned URLs:

Wikipedia,wikipalapp,wiktionary,facebook,weather,.gov.,panorama,cottage,hotel,for-sale,airbnb,booking,expedia,bustimes,books,pubguide, citypopulation,laterooms,indeed.co.uk, youtube,inn,rentals,yr.no,ordnancesurvey,bedandbreakfast,news,bbc,transportation,streetviewmaps,campsites,nurseries,mypub,forecast,prices,selfcatering,home,countrysideclassroom,geog.port.ac.uk,lakelandcampingbarns,fortune,money,business, observer,shop,cafe,scandal,store,twitter,forbes,linkedin,theguardian,finestproperties,publications,naturalengland,distillery,restaurant,dictionary, highclose,house,availability,onthemarket,windguru,colinday,rooms,aqua3,ferries,property,wikimedia,delivery,ancestry,holidaylettings, accommodation,jobs,online,brew,House,Hotel,tate.org.uk,playdale,finance,solutions,amazon,imdb,thetimes,nytimes,telegraph.co.uk,independent,



Appendix 1. Temporal distribution of collected texts. Left: locations of the search terms from the 'TripAdvisor list'. Right: locations of the search terms from the 'Wainwright list'.

Table Appendix 1
Rules and important distinguishing marks for annotation of first-person perception training data.

First-person landscape descriptions	Not relevant descriptions
• explicit descriptions of perception (Heard Snipe piping for the first time this year; the heather smells lovely; the scent of wet peat and sun-warmed bog myrtle) • events that have already happened as opposed to anticipated ones • descriptions using verbs of motion in combination with personal pronouns 'I' or 'we' (we went to …; I walked 12 miles) • potentially contain references to time (today; Wednesday; this lovely spring morning) • potentially contain descriptions of weather (it was still raining; the sun was shining) • potentially include names of the fellow travelers	• describe anticipated events (next week we go to the magnificent Aira Force waterfall) • present a consistent use of passive voice (it can be done by both car and on foot) • contain imperatives (keep on the road, head north) • contain information to help navigation (you reach; the river on your right) • include lists of walks (five best walks in the Lake District) • are indoors descriptions • weather forecasts • official parish information

Table Appendix 2
Taxonomy of tranquillity for annotation.

Classes of tranquillity	Description
Combination of sight and sound perception	Descriptions, where visual attributes of the scene are as important as sound, and where absence of sounds is implicit (e.g., 'A remembrance service is held here every year and I can't think of a more beautiful and peaceful place to reflect.')[1]
Contrasting sounds	Descriptions reflecting ephemerality of tranquillity by comparing it to other less tranquil locations, different time of day or mentioning sounds which add or detract from overall tranquillity (e.g., 'A moment of peace at Ashness Bridge – rare moments indeed!'[2])
No-movement	Explicit mention of a lack of movement with implied silence and tranquillity (e.g., 'Below to the west Buttermere appeared mirror calm, the blue of the sky reflected deeply in its chill waters.'[3])
Total silence and tranquil sounds	Either descriptions of tranquil sounds without contrast or explicit descriptions of complete silence (e.g., 'Further to the north Blencathra and Skiddaw put in an appearance in the evening sun, and we stopped to listen to the silence – not a sound – very peaceful and relaxing.'[4])

[1] http://juliahedges.blogspot.com/2018/06/a-walk-up-mighty-great-gable.html.
[2] http://www.lakedistrict-walks.co.uk/2016/September/10.09.2016_Bleaberry_Fell.html.
[3] http://www.david-forster.com/section278539_221484.html.
[4] http://www.ramblingpete.walkingplaces.co.uk/day/lakes/martindale.htm.

filmclub,realestate,resort,medical,trains,yellowpages,checkmypostcode,geograph.org.uk,britishplacenames,office,fivestar,research,quora.com,product,clinic,agency,police,science,street,ebay,etsy,paranormal,massage,pinterest,.fr,.au,.jp,.edu,timeanddate,obituaries,lyrics,stackexchange,dailypost,lawyers,washingtonpost,glamour,movies,usatoday,.cnn.,denverpost,metro.co.uk,thesun.co.uk,wiki,scientificamerican,person,startribune,religion,tvtropes,resetera,soundcloud,rottentomatoes,itunes,nypost,britannica,salvationarmy,psychologytoday,cancer,edinburghfestival,vancouversun,spotify,goodreads,foxsports,nbcsports,seattletimes,encyclopedia,dailystar,biography,tvguide,zhidao.baidu,flickr,accident,incident,incidents,cars,gamespot,whitepages,vimeo,startrek,lodging-world,the-saleroom,sun-up.co.uk

Keywords sound:

Babble,bang,beat,beep,blare,blast,boom,bubble,burble,burr,chime,chink,chir,chug,clack,clang,clank,clap,clash,clatter,cling,clink,clomp,clump,clunk,crack,crackle,crash,creak,crepitate,crunch,cry,ding,dong,explode,fizz,fizzle,groan,gurgle,hum,jangle,jingle,knell,lilt,moan,murmur,patter,peal,ping,plink,plonk,plop,plunk,pop,putter,rap,rasp,rattle,ring,roll,rumble,rustle,shriek,shrill,sizzle,splash,splutter,sputter,squelch,strike,swish,swoosh,thrum,thud,thump,thunder,thunk,tick,ting,tinkle,toll,toot,tootle,-trumpet,twang,ululate,vroom,wheeze,whine,whir,whish,whoosh,whump,zing,baa,bark,bay,bellow,blat,bleat,bray,buzz,cackle,call,caw,chatter,cheep,chirp,chirrup,chitter,cluck,coo,croak,crow,cuckoo,drone,gobble,growl,grunt,hee,haw,hiss,honk,hoot,howl,meow,mew,moo,neigh,oink,peep,pipe,purr,quack,roar,scrawk,scream,screech,sing,snap,snarl,snort,snuffle,squawk,squeak,squeal,stridulate,trill,tweet,wail,warble,whimper,whinny,whistle,woof,yap,yell,yelp,yip,yowl,din,echo,resonate,resound,sound,listen,hear,clamor,shout,holler,noise,brawl,discord,grate,gnash,grind,slam,stamp,surd,clonk,blether,blither,ripple,guggle,brattle,chirr,clangor,grumble,whoop,boisterous,shrill,silent,melodic,clamorous,melodious,croaky,muffled,soundless,discordant,squeaky,noiseless,earsplitting,noisy,tacit,thundering,gruff,thunderous,quiet,rasping,tuneful,raspy,raucous,resonant,vociferous,hoarse,rowdy,husky,creaky,loud,screaming,screechy,deafening,hushed,inaudible,stillness,muteness,still,silentness,soundlessness,noiselessness,flick,whisper,susurration,mutter,tranquillity,tranquility,peace,restfulness,quietness,calm,calmness,quietude,serenity,peacefulness,reposefulness,shush,whisht,whiffle,mute,tacitly,quietlike,mouselike,tongueless,dumb,mousy,whisperless,voiceless,halcyon,peaceful,peaceable,restful,tranquil,undisturbed

Keywords smells:

Reek,smell,stink,inhale,aroma,odour,scent,malodor,malodour,stench,fetor,mephitis,acridity,aroma,fragrance,acrid,antiseptic,foetid,fetid,fragrant,musky,musty,noxious,whiffy,odorous,pungent,putrid,rancid,scentless

Scenic pairs:

('middle','distance'),('far','distance'),('steep','slopes'),('highest','point'),('lower','slopes'),('low','tide'),('long','ridge'),('small','loch'),('fine','view'),('northern','slopes'),('distant','view'),('small','lochan'),('high','point'),('south','ridge'),('small','cairn'),('western','slopes'),('west','ridge'),('small','island'),('southern','slopes'),('prominent','hill'),('south','shore'),('rough','grazing'),('east','ridge'),('far','side'),('small','islands'),('distant','hill'),('steep','slope'),('high','tide'),('natural','arch'),('surrounding','hills'),('southern','side'),('south','coast'),('eastern','slopes'),('rocky','outcrops'),('eastern','side'),('small','hill'),('prominent','peak'),('small','crags'),('freshwater','loch'),('northern','shore'),('steep','ridge'),('eastern','top'),('fine','views'),('superb','views'),('highest','-hill'),('large','boulder'),('great','views'),('good','views'),('good','view'),('gentle','slopes'),('sharp','peak'),('moderate','slopes'),('rough','grass'),('remote','-hill'),('coastal','scenery'),('southern','ridge'),('conical','hill'),('narrow','ridge'),('deep','gorge'),('broad','ridge'),('near','distance'),('west','coast'),

('rough','moorland'),('unnamed','lochan'),('lower','part'),('magnificent','views'),('good','path'),('cliff','top'),('south','end'),('upper','part'),('great','view'),('flat','summit'),('distant','views'),('similar','view'),('steep','valley'),('rocky','spur'),('rough','track'),('true','summit'),('highest','peak'),('rocky','promontory'),('lower','top'),('old','pier'),('fresh','snow'),('rocky','hill'),('lewisian','gneiss'),('huge','boulder'),('main','summit'),('exposed','rock'),('west','shore'),('rocky','summit'),('beautiful','beach'),('northwest','ridge'),('big','hill'),('northern','side'),('southwest','ridge'),('fine','viewpoint'),('cliff','edge'),('rough','slopes'),('northern','ridge'),('small','lochans'),('little','hill'),('low','point'),('south','corner'),('western','side'),('brown','trout'),('northern','tip'),('far','shore'),('high','ground'),('shallow','loch'),('eastern','shore'),('southern','tip'),('tussocky','grass'),('small','bay'),('coastal','path'),('small','waterfall'),('southern','shore'),('steep','descent'),('sandy','beach'),('few','places'),('stepping','stones'),('many','summits'),('more','rocks'),('rocky','pavement'),('rocky','bit'),('shapely','peak'),('cambrian','quartzite'),('sharp','ridge'),('topped','mountain'),('small','trout'),('horizontal','strata'),('early','snow'),('low','hill'),('north','slopes'),('dissected','bog'),('long','loch'),('largest','loch'),('highest','mountain'),('beautiful','scenery'),('superb','view'),('top','end'),('rocky','slopes'),('tiny','lochan'),('made','path'),('eastern','part'),('deep','pool'),('good','viewpoint'),('rocky','coastline'),('unnamed','top'),('tidal','island'),('little','snow'),('rough','hill'),('rocky','coast'),('steep','ascent'),('wonderful','view'),('wonderful','views'),('west','face'),('small','hills'),('lowest','point'),('upper','valley'),('covered','slopes'),('southeast','side'),('spectacular','view'),('north','end'),('clear','view'),('south','top'),('small','outcrop'),('steep','drop'),('stunning','views'),('unnamed','hill'),('rocky','beach'),('rocky','ridge'),('deep','snow'),('right','side'),('shaped','valley'),('right','skyline'),('heavy','snow'),('left','skyline'),('boggy','ground'),('glen','floor'),('gentle','ridge'),('flat','floor'),('facing','slopes'),('faint','path'),('far','horizon')

Unattractive pairs:

('industrial','estate'),('dual','carriageway'),('new','housing'),('new','development'),('small','estate'),('residential','area'),('new','estate'),('industrial','units'),('terraced','houses'),('slip','road'),('former','airfield'),('new','building'),('large','estate'),('cooling','towers'),('many','buildings'),('detached','houses'),('modern','estate'),('busy','junction'),('modern','housing'),('new','centre'),('retail','park'),('new','developments'),('modern','building'),('small','development'),('new','station'),('former','factory'),('recent','development'),('main','runway'),('industrial','area'),('northbound','carriageway'),('new','buildings'),('large','complex'),('former','village'),('suburban','road'),('high','voltage'),('multi','storey'),('new','park'),('major','road'),('terraced','housing'),('major','junction'),('central','reservation'),('wartime','airfield'),('local','shops'),('residential','road'),('detached','housing'),('former','garage'),('filling','station'),('large','park'),('retail','outlet'),('large','village'),('large','roundabout'),('industrial','unit'),('new','units'),('large','development'),('large','centre'),('typical','housing'),('closed','pub'),('new','part'),('small','businesses'),('local','centre'),('perimeter','fence'),('new','stadium'),('new','estates'),('single','carriageway'),('adjacent','site'),('carriageway','road'),('new','homes'),('new','block'),('new','turbines'),('new','roundabout'),('current','building'),('old','a1'),('overflow','park'),('main','office'),('back','street'),('social','club'),('chinese','takeaway'),('flat','roofs'),('industrial','premises'),('main','station'),('new','apartments'),('more','housing'),('old','machinery'),('named','road'),('residential','street'),('solar','panels'),('private','houses'),('local','road'),('special','train'),('general','store'),('typical','houses'),('public','houses'),('near','junction'),('local','office'),('old','works'),('staggered','junction'),('new','lease'),('old','base'),('heavy','industry'),('modern','style'),('industrial','use'),('heavy','plant'),('new','store'),('slip','roads'),('new','blocks'),('main','stand'),('new','hall'),('main','hall'),('typical','development'),('several','streets'),('busy','roundabout'),('many','businesses'),('high','tension'),('suburban','street'),('new','premises'),('main','carriageway'),('large','works'),('agricultural','produce'),('crossing','gates'),('organic','farm'),('parked','cars'),('industrial','site'),('much','traffic'),('underground','reservoir'),('overhead','lines'),('new','construction'),('old','factory'),('multiple','unit'),('rubbish','tip'),('terraced','cottages'),('adjacent','park'),('residential','areas'),('static','caravans'),('old','hospital'),('overhead','cables'),('new','hospital'),('main','centre'),('new','extension'),('staggered','crossroads'),('major','development'),('unusual','design'),('next','station'),('large','sheds'),('large','hospital'),('busy','intersection'),('free','house'),('many','estates'),('residential','estate'),('new','facilities'),('wartime','factory'),('large','tower'),('underground','workings'),('former','depot'),('large','plant'),('new','flats'),('concrete','building'),('motorway','junction'),('current','station'),('large','barns'),('industrial','complex'),('striking','building'),('main','industry'),('modern','centre'),('local','service'),('guided','busway'),('multi','purpose'),('tall','buildings'),('double','glazing'),('suburban','housing'),('light','units'),('high','density'),('many','stations'),('retail','centre'),('nearby','line'),('largest','centre'),('agricultural','equipment'),('executive','houses'),('main','hospital'),('electric','pumps'),('mini','roundabout'),('closed','station'),('bound','carriageway'),('new','build'),('other','facilities'),('large','quarry'),('substantial','buildings'),('industrial','estates'),('built','houses'),('busy','carriageway'),('rolling','stock'),('large','blocks'),('former','estate'),('small','site'),('last','building'),('front','wall'),('agricultural','machinery'),('big','houses'),('large','buildings'),('suburban','development'),('main','gates'),('former','works'),('old','runways'),('built','building'),('large','range'),('old','runway'),('more','buildings'),('rural','lane'),('integral','part'),('third','rail'),('second','bridge'),('tidy','farm'),('arterial','road'),('main','lines'),('mobile','homes')

List of British mammals:

Beaver,vole,mouse,rat,dormouse,squirrel,porcupine,hare,rabbit,mole,shrew,hedgehog,bat,pipistrelle,dog,fox,seal,walrus,marten,weasel,polecat,otter,badger,wildcat,cat,mink,coati,boar,goat,sheep,cattle,deer,reindeer,moose,muntjac,buffalo,whale,dolphin,beluga,porpoise,orca,cow,stag,cattle,lamb

List of natural elements:

Water,river,tree,beach,sea,snow,coast,stone,rain,grass,harbour,seaside,leaves,lake,wood,plant,sand,pond,mist,fog,ice,rock,forest,hill,island,leaf,mountain,bay,waterfall,loch,wave,seafront,mud,landscape,summit,valley

Annotated data of Geograph: https://github.com/olgaches/Geograph_sound_descriptions.

References

Arias, F. J. C. (2019). Fuzzy String Matching in Python [WWW Document]. URL: https://www.datacamp.com/community/tutorials/fuzzy-string-python.

Barnard, J., Jackson, G., & Bloomer, J. (n.d.) England has a new mountain: Miller Moss. Now go find it [WWW Document]. 2018. URL: https://www.grough.co.uk/magazine/2018/08/09/england-has-a-new-mountain-miller-moss-now-go-find-it (accessed 06.10.19).

Bieling, C. (2014). Cultural ecosystem services as revealed through short stories from residents of the Swabian Alb (Germany). *Ecosystem Services, 8*, 207–215.

Bieling, C., Plieninger, T., Pirker, H., & Vogl, C. R. (2014). Linkages between landscapes and human well-being: An empirical exploration with short interviews. *Ecological Economics, 105*, 19–30.

Bing Web Search [WWW Document]. (2019). URL: https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/ (accessed 10.02.19).

boyd, d. (2007). Why youth (heart) Social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.). *MacArthur foundation series on digital learning – Youth, identity, and digital media volume*. Cambridge, MA: MIT Press.

Brown, G., & Reed, P. (2009). Public participation GIS: A new method for use in national forest planning. *Forest Science, 55*, 166–182.

Bruns, D., & Stemmer, B. (2018). Landscape assessment in Germany. *Routledge handbook of landscape character assessment* (pp. 154–167). .

Butler, A. (2016). Dynamics of integrating landscape values in landscape character assessment: The hidden dominance of the objective outsider. *Landscape Research, 41*, 239–252.

Carles, J. L., Barrio, I. L., & De Lucio, J. V. (1999). Sound influence on landscape values. *Landscape and Urban Planning, 43*, 191–200.

Caspersen, O. H. (2009). Public participation in strengthening cultural heritage: The role of landscape character assessment in Denmark. Geogr. Tidsskr. – Danish. *Journal of Geography, 109*, 33–45.

Chesnokova, O., & Purves, R. S. (2018a). *Automatically creating a spatially referenced corpus*

of landscape perception. 12th ACM SIGSPATIAL workshop on geographic information retrieval. Seattle, WA, USA: ACM.
Chesnokova, O., & Purves, R. S. (2018b). From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text. *Applied Geography, 93*, 103–111.
Chesnokova, O., Taylor, J. E., Gregory, I. N., & Purves, R. S. (2019). Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District. *International Journal of Geographical Information Science, 33*, 2430–2454.
Clemetsen, M., Krogh, E., & Thorén, K. H. (2011). Landscape perception through participation: Developing new tools for landscape analysis in local planning processes in Norway. In M. Jones, & M. Stenseke (Eds.). *The European landscape convention. Challenges of participation* (pp. 219–237). Springer.
Coates, P. A. (2005). The strange stillness of the past: Toward an environmental history of sound and noise. *Environmental History Durh. N. C. 10*, 636–665.
Council of Europe. (2000). European landscape convention. Rep. Conv. Florence ETS No. 17, 8.
Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision, 7*, 81–227.
Cronon, W. (1992). A place for stories: Nature, history, and narrative. *Journal of American History, 78*, 1347–1376.
Dann, G. M. S., & Jacobsen, J. K. S. (2003). Tourism smellscapes. *Tourism Geography, 5*, 3–25.
Dara-Abrams, D. (2011). Jenks natural breaks [WWW Document]. URL: https://gist.github.com/drewda/1299198.
Daume, S., Albert, M., & von Gadow, K. (2014). Forest monitoring and social media – Complementary data sources for ecosystem surveillance? *Forest Ecology and Management, 316*, 9–20.
Davies, C. (2013). Reading geography between the lines: Extracting local place knowledge from text. *Lecture Notes in Computer Science (including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics), 8116 LNCS*, 320–337.
de Kunder, M. (2019). The size of the World Wide Web (The Internet) [WWW Document]. URL: https://www.worldwidewebsize.com/ (accessed 10.01.19).
Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., ... Danforth, C. M. (2014). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences, 112*, 2389–2394.
Donaldson, C., Gregory, I. N., & Taylor, J. E. (2017). Locating the beautiful, picturesque, sublime and majestic: Spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography, 56*, 43–60.
Edwards, J. (2018). Literature and sense of place in UK landscape strategy. *Landscape Research, 44*, 659–670.
Gonzalez, J., Rodrigues, P., & Cohen, A. (2017). Fuzzywuzzy: Fuzzy string matching in python [WWW Document]. URL: https://github.com/seatgeek/fuzzywuzzy.
Esri. (2019). ArcGIS API for Python [WWW Document]. URL: https://pro.arcgis.com/en/pro-app/arcpy/get-started/arcgis-api-for-python.htm.
Fairclough, G., Sarlöv Herlin, I., & Swanwick, C. (Eds.). (2018). *Routledge handbook of landscape character assessment*. Routledge.
Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MA: MIT Press.
Fisher, J. A. (1999). The value of natural sounds. *Journal of Aesthetic Education, 33*, 26–42.
Galaz, V., Crona, B., Daw, T., Bodin, Ö., Nyström, M., & Olsson, P. (2010). Can web crawlers revolutionize ecological monitoring? *Frontiers in Ecology and the Environment, 8*, 99–104.
Granö, J. G. (1997). *Pure geography*. John Hopkins University Press.
Greenaway, M. (2017). Web-scraping policy [WWW Document]. URL: https://www.ons.gov.uk/aboutus/transparencyandgovernance/lookingafterandusingdataforpublicbenefit/policies/policieswebscrapingpolicy (accessed 07.01.19).
Hale, S. A., Blank, G., & Alexander, V. D. (2017). Live versus archive: Comparing a web archive and to a population of webpages. In N. Brügger, & R. Schroeder (Eds.). *The web as history* (pp. 45–61). London: UCL Press.
Herlin, I. S. (2016). Exploring the national contexts and cultural ideas that preceded the Landscape Character Assessment method in England. *Landsc. Res. 41*, 175–185.
Hewlett, D., Harding, L., Munro, T., Terradillos, A., & Wilkinson, K. (2017). Broadly engaging with tranquillity in protected landscapes: A matter of perspective identified in GIS. *Landsc. Urban Plann. 158*, 185–201.
Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing 1373–1378.
Jockers, M. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
Joho, H., & Sanderson, M. (2000). Retrieving descriptive phrases from large amounts of free text. *Proceedings of the ninth international conference on information and knowledge management (CIKM)* (pp. 180–186). USA: Mclean.
Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science, 22*, 1045–1065.
Jones, M., & Stenseke, M. (Eds.). (2011). *The European landscape convention. Challenges of participation*. Springer.
Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. *EMNLP-CoNLL. Prague* (pp. 1075–1083). .
Kienast, F., Degenhardt, B., Weilenmann, B., Wäger, Y., & Buchecker, M. (2012). GIS-assisted mapping of landscape suitability for nearby recreation. *Landscape and Urban Planning, 105*, 385–399.
Kienast, F., Frick, J., van Strien, M. J., & Hunziker, M. (2015). The Swiss landscape monitoring program – A comprehensive indicator set to measure landscape change. *Ecological Modelling, 295*, 136–150.
Kienast, F., Wartmann, F., Zaugg, A., & Hunziker, M. (2019). *A Review of Integrated Approaches for Landscape Monitoring*.
Kolen, J., & Renes, J. (2015). Landscape biographies: Key issues. In J. Kolen, H. Renes, & R. Hermans (Eds.). *Landscape biographies* (pp. 21–48). Amsterdam University Press.
Kolen, J., Renes, H., & Bosma, K. (2018). The landscape biography approach to landscape characterisation. *Routledge handbook of landscape character assessment* (pp. 168–184). .
Komossa, F., van der Zanden, E. H., Schulp, C. J. E., & Verburg, P. H. (2018). Mapping landscape potential for outdoor recreation using different archetypical recreation user groups in the European Union. *Ecological Indicators, 85*, 105–116.
Krause, B. (2008). Anatomy of the soundscape. *Journal of the Audio Engineering Society*, 56.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
Lawson, R. (2015). *Web scraping with Python*. Packt Publishing Ltd.
Leidner, J. L., Sinclair, G., & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. *Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references. Stroudsburg, PA, USA* (pp. 31–38). .
Levin, B. (1993). *English verb classes and alternations*. University of Chicago Press.
Liu, B. (2012). *Sentiment analysis and opinion mining*. Toronto: Morgan & Claypool Publishers.
Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach. *WWW 2011 – Session: semantic analysis. Hyderabad, India* (pp. 347–356). .
Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods, 41*, 558–564.
MacFarlane, R., Haggett, C., Fuller, D., Dunsford, H., & Carlisle, B. (2004). Tranquillity mapping: Developing a robust methodology for planning support.
Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition, 130*, 266–270.
Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. The MIT Press.
Mark, D. M., & Turk, A. G. (2017). Ethnophysiography. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, 1–11.
Moncla, L., Gaio, M., & Mustière, S. (2014). Automatic itinerary reconstruction from texts. *Eighth international conference on geographic information science (GIScience 2014)* (pp. 253–267). .
Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014). *Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems. Dallas/Fort Worth, TX, USA*.
Nielsen, J. (2006). The 90-9-1 rule for participation inequality in social media and online communities [WWW Document]. URL: http://www.useit.com/alertbox/participation_inequality.html.
Nomination. (n.d.) Nomination of the English Lake District for inscription on the world heritage list. [WWW Document]. 2017. URL: https://whc.unesco.org/en/list/422 (accessed 07.09.19).
Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science, 22*, 265–287.
Palmer, C., & Brady, E. (2007). Landscape and value in the work of Alfred Wainwright (1907–1991). *Landscape Research, 32*, 397–421.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Blondel, M. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825–2830.
Pheasant, R., Horoshenkov, K., Watts, G., & Barrett, B. (2008). The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places? *Journal of the Acoustical Society of America, 123*, 1446–1457.
Pheasant, R. J., & Watts, G. R. (2015). Towards predicting wildness in the United Kingdom. *Landscape and Urban Planning, 133*, 87–97.
Prior, J. (2017). Sonic environmental aesthetics and landscape research. *Landscape Research, 42*, 6–17.
PostGIS. (2019). DBScan Clustering [WWW Document]. URL: https://postgis.net/docs/ST_ClusterDBSCAN.html.
Quercia, D., & Schifanella, R. (2015). *Smelly maps: The digital life of urban smellscapes. 9th international AAAI conference on web and social media. Oxford, UK*.
Rattenbury, T., & Naaman, M. (2009). Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web, 3*, 1–30.
Relph, E. (1976). *Place and placelessness*. London: Pion Press.
Richards, D. R., & Tunçer, B. (2018). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem Services, 31*, 318–325.
San Roque, L., Kendrick, K. H., Norcliffe, E., Brown, P., Defina, R., Dingemanse, M., ... Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics, 26*, 31–60.
Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering, 8*, 970–974.
Srinivasa-Desikan, B. (2018). *Natural language processing and computational linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.
Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *5th international workshop on semantic evaluation* (pp. 321–324). Uppsala, Sweden: ACL. Association for Computational Linguistics.
Swanwick, C., & Fairclough, G. (2018). Landscape character: Experience from Britain. In G. Fairclough, I. Sarlöv Herlin, & C. Swanwick (Eds.). *Routledge handbook of landscape character assessment* (pp. 21–36). Routledge.

Taylor, J. E. (2018). Echoes in the mountains: The romantic lake district's soundscape. *Studies in Romanticism, 57*, 383–406.
Taylor, J. G., Czarnowski, K. J., & Flick, S. (1995). The importance of water to Rocky Mountain National Park visitors: An adaptation of visitor-employed photography to natural resources management. *Journal of Applied Recreation Research, 20*, 61–85.
Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., ... Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation, 233*, 298–315.
Tudor, C. (2014). An approach to landscape character assessment.
van den Bosch, A., Bogers, T., & de Kunder, M. (2016). Estimating search engine index size variability: A 9-year longitudinal study. *Scientometrics, 107*, 839–856.
Wartmann, F. M., Acheson, E., & Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: An interdisciplinary approach. *International Journal of Geographical Information Science, 32*, 1–21.
Watkins, D. (2008). Landscape character assessment and guidelines.
Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition, 179*, 213–220.
Zachara, M., & Palka, D. (2016). Comparison of text-similarity metrics for the purpose of identifying identical web pages during automated web application testing. *Information systems architecture and technology: Proceedings of 36th international conference on information systems architecture and technology – ISAT 2015 – Part II* (pp. 25–35). .
Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media Society,* 4.

LIST OF PRESENTATIONS, PUBLICATIONS AND WORKSHOPS

Presentations since 2016

07/2019	**IALE, Milan**
	Extracting perceived landscape properties from texts for Landscape Character Assessment
11/2018	**SIGSPATIAL, GIR workshop, Seattle**
	Automatically creating a spatially referenced corpus of landscape perception
06/2018	**GeoSummit, Bern**
	Extracting landscape properties from text | Poster
09/2017	**IALE, Gent**
	Landscape preference assessment from digital sources: Comparing historical texts and annotated images
09/2017	**COSIT, L'Aquila**
	A crowdsourced model of landscape preference
09/2017	**COSIT, L'Aquila**
	Lake District Soundscapes: Analysing aural experience through text | Poster
09/2016	**GIScience, Montréal**
	Comparing digital traces of modern travellers to journeys of two 18th-19th century British poets
09/2016	**Spatial Humanities, Lancaster**
	Landscape descriptions in historical and modern landscape corpora
03/2016	**Esri Campus Day, Wädenswil**
	To what extent can aesthetic values of landscapes be extracted from modern diaries of hikers?

Peer-reviewed publications since 2016

Koblet, O. and Purves, R.S., 2020. From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally. Landscape and Urban Planning, Volume 197, 103757.

Bahrehdar A.R., **Koblet, O.** and Purves, R.S., 2019. Approaching location-based services from a place-based perspective: from data to services?, Journal of Location Based Services, 13:2, 73-93.

Chesnokova, O., Taylor, J.E., Gregory, I.N., and Purves, R.S., 2019. Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District. International Journal of Geographical Information Science, 33:12, 2430-2454.

Chesnokova, O. and Purves, R.S., 2018. Automatically creating a spatially referenced corpus of landscape perception. 12th ACM SIGSPATIAL Workshop on Geographic Information Retrieval, Seatlle, USA.

Chesnokova, O. and Purves, R.S., 2018. From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text. Applied Geography, 93, 103–111.

Chesnokova, O., Nowak, M., and Purves, R.S., 2017. A crowdsourced model of landscape preference. In: E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore, eds. 13th International Conference on Spatial Information Theory (COSIT 2017). Leibniz International Proceedings in Informatics, 1–13.

Chesnokova, O., Gregory, I.N., and Purves, R.S., 2016. Comparing digital traces of modern travellers to journeys of two 18th-19th century British poets. International Conference on GIScience Short Paper Proceedings, 1(1):45-48.

Co-organised workshops

04/2019	**Environmental Narratives**, Independent workshop, with Ross Purves and Ben Adams
03/2018	**Who is behind your data? A conversation across geographic disciplines**, InnoPool, with GIUZ members
09/2017	**Spatial Humanities meets Spatial Information Theory: Place, Space, and Time in Humanities Research**, COSIT Workshop, with Ben Adams and Karl Grossner