Ann-Marie Eklund

The game of health search

Data linguistica

<http://www.svenska.gu.se/publikationer/data-linguistica/>

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

26 • 2014

Ann-Marie Eklund

The game of health search

Gothenburg 2014

Data linguistica 26
ISBN 978-91-87850-55-4
ISSN 0347-948X

Printed in Sweden by
Ineko AB Göteborg 2014

Typeset in LyX by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
The game of health search
by Magnus Andersson ©

Author photo on back cover by Magnus Andersson

ABSTRACT

Almost two of three Swedes use internet to search for health related information on diseases, treatments and care givers. This is in line with the stated public goals to establish a digital complement to the traditional doctor's visits and calls to health centres for medical advice. Moreover, mobile devices such as smartphones and tablets are increasingly used to carry out these activities, and it raises the question on how a health information portal should behave to support the needs of today's and tomorrow's information seekers.

In this work we present, to our knowledge, the first analysis of the use of the official health information portals 1177.se and vardguiden.se with a focus on describing the relations between seekers and portals, as expressed by the language of queries and answers. Of special interest is the role of the language as a means to establish and maintain the seekers' trust in a portal as a complement to doctor's visits and calls. As a result of our efforts, we are able to present a number of principles of behaviour to which we believe a portal should adhere to be trustworthy in the eyes of the seekers.

We also introduce a conceptual framework with a basis in game-theoretic models of rational behaviour, and the use of error analysis of second-language learning and stylistics studies of written texts, to provide a setting for descriptive and predictive analysis of information search as an interaction between actors comprising seekers and portals.

SAMMANFATTNING

Nära två av tre svenskar använder internet för att söka efter hälsorelaterad information om till exempel sjukdomar, behandlingar och vårdgivare. Detta är i linje med samhällets mål att etablera ett digitalt komplement till traditionella läkarbesök och telefonsamtal till vårdcentraler för att få medicinska råd. Dessutom ökar användningen av mobila enheter som smartphones och surfplattor för hälsosökning och det väcker frågan hur en internetportal för hälsoinformation bör vara utformad för att stödja dagens och morgondagens informationssökares behov.

I detta arbete presenteras, såvitt vi vet, den första studien av användningen av de officiella hälsoinformationsportalerna 1177.se och vardguiden.se med fokus på att beskriva interaktionen mellan sökare och portal i form av språket som används i frågor och svar. Särskilt intressant är språkets roll som ett sätt att erhålla och vidmakthålla sökarnas förtroende för portalen som ett komplement till läkarbesök och telefonsamtal. Som ett resultat av vår studie presenterar vi ett antal principer som vi anser att portaler bör följa för att ge ett trovärdigt intryck.

Vi introducerar även ett konceptuellt ramverk, med en grund i spelteoretiska modeller av rationellt beteende, felanalys inom andraspråksinlärning och stilistik för skriven text, för att erbjuda en grund för deskriptiv och prediktiv analys av informationssökning som en interaktion mellan aktörer i form av sökare och portaler.

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisors Dimitris Kokkinakis, Jussi Karlgren and Lars Borin for their support and valuable comments during the writing of this thesis and Hercules Dalianis for reviewing an earlier version.

I would like to thank the Graduate school of language technology (GSLT), Centre for language technology (CLT), Wilhelm och Martina Lundgrens Vetenskapsfond 1 and Filosofiska fakulteternas gemensamma donationsnämnd for their financial support.

Without Svetoslav Marinov at Findwise and Euroling AB in agreement with Stockholm County Council (Jessika Bjurel) providing the search logs this work would not have been possible.

This thesis is partly a result of collaborations with Dimitris Kokkinakis, Svetoslav Marinov, Farnaz Moradi, Daniela Oelke, Tomas Olovsson and Philippas Tsigas, and an unknown number of reviewers.

Thanks to Dana Dannélls and Kristina Holmlid for helping out with typographic and layout challenges.

I also want to thank colleagues and friends at Centre for Language Technology, the HEXAnord network (HEalth TeXt Analysis network in the Nordic and Baltic countries), Department of Swedish and Språkbanken.

Finally, I want to thank Magnus for his support over the years.

CONTENTS

PROLOGUE

> An evening in August at her summerhouse in Norrtälje, Julia suffers from a stiff neck and fever. Using her smartphone she searches for information at vardguiden.se by posting the query *stel i nacken feber* 'stiff in the neck fever'. Among the 20 first answers[1], the majority concerns topics like measles and fever in children. However, in the small gists[2] describing the answers she spots one titled TBE[3] mentioning *fästing* 'tick' which triggers her interest. She clicks the hyperlink to find out more, especially that her problems may be signs of encephalitis requiring immediate treatment. Since the portal is regulated by the government, she trusts the information and decides to visit the closest emergency unit for a possible confirmation that she suffers from TBE.

According to the Swedish government, the aims of the official internet health portals 1177.se and vardguiden.se are to promote health and empower the citizens by providing reliable and easily accessible online health information, and facilitate the health care process and contacts with care givers. For instance, in January 2013 the Stockholm County Council reported that vardguiden.se had 2 million visitors per month.

Julia's interaction was registered, or logged, by vardguiden.se, and below is an example of the type of registered information. In addition to the query *stel i nacken feber* the search log contains the explicit information, or context, defining that the search was carried out using a mobile device, in Norrtälje in August and that the seeker chose the 12th of 20 presented answers.[4]

[1] An *answer* is a collection of information describing a topic, possibly in the form of a web page with an associated *hyperlink*, i.e. clickable reference, pointing to it.

[2] An *answer gist* is a small paragraph of text describing the essence of an answer, possibly also with a hyperlink to the answer.

[3] "Tick-borne encephalitis (TBE) is a viral infectious disease involving the central nervous system. The disease is mostly manifested as meningitis, encephalitis or meningoencephalitis. [...] The tick-borne encephalitis virus is known to infect a range of hosts including ruminants, birds, rodents, carnivores, horses and humans." (Wikipedia 2013: Tick-borne encephalitis)

[4] This example is hypothetical, but reflects the type of queries posted and the registered information. Furthermore, the numbers correspond to the ones obtained if the query would have been posted at the given date and location.

```
2013-08-21:18-43-05, Norrtälje, Mobile,
stel i nacken feber, 12, 20
```

If we consider the purpose of vardguiden.se, we may ask ourselves if the portal could have done better in predicting the answers of interest to Julia. The context did reference both Norrtälje and August, and this area is known to be prone to ticks during summer. Moreover, if she would have posted the query *nackstelhet feber* 'neck stiffness fever', with the same meaning as the original one, the TBE answer would have occurred at the top of the list of answers. It is also worth mentioning that it was just by chance she noticed the reference to ticks that triggered her interest, and if this word would not have been mentioned in the gist it is possible that she would not have chosen the answer implicating the need of immediate care.[5]

As an information provider on health topics it is important to try to predict the seeker's needs as well as possible, especially when she might be in distress and adequate information be the dividing line between suffering and care. Let us use Julia's interaction as a starting point for a brief reflection on how an information provider could utilise the query and its context, in addition to the rest of the logged interactions and the portal's knowledge base, to provide better and trustworthy support to the public.

There are several different ways to achieve an improved interaction between an information seeker and a portal, and to facilitate the presentation we make use of the observation that the communication between them can be viewed as a "game" where the seeker makes the first move by posting a query, possibly trying to foresee how the portal will react. The portal tries to understand the seeker by analysing the query and the provided context, ending up with different interpretations. Each interpretation will then lead to the portal providing a corresponding, possibly empty, list of answers to the seeker – the move of the portal. The seeker, who after querying was in a situation expecting a certain type of answers, then considers the portal's move and decides on whether to continue with a next move by posting a new or refined query, to be satisfied with the answer and finish the interaction or leave the game unsatisfied. The degree of seeker satisfaction may also impact her trust in the portal. For instance, no answers or many "irrelevant" answers to a query which seems trivial to the seeker, may result in distrust and decreased use. The unfolding of the interaction can be viewed as a path of queries, answers and situations,

[5] As of November 2013 the two portals 1177.se and vardguiden.se have been merged into one, and in the new setting the reference to ticks in the gist has disappeared, the query *nackstelhet feber* 'neck stiffness fever' only leading to three answers in comparison to 30 for its semantic equivalent *stel i nacken feber* 'stiff in the neck fever' (2014-03-28). However, even worse is that the new portal does not even provide TBE as a specific answer.

or as a sequence of (potential) moves with different payoffs, i.e. the expected gain for an actor by her move. In our setting, the level of payoff for a seeker indicates how happy she is with an answer, and for the portal it reflects the degree of certainty that the answer was the one asked for by a seeker.

In the case of Julia, she was possibly in a situation where she was interested in treatments of her symptoms, and probably did not want to learn more about TBE vaccination. Thereby, after posting her query she was expecting the portal to "understand" this need and answer accordingly. Hence, Julia would consider to gain more by answers on TBE treatment than prevention and the former would have higher payoff to her. However, the portal had access only to the query and the context, and depending on its ability to interpret these could come to different conclusions regarding Julia's need. That is, the portal expected that it would gain more by providing prevention answers in comparison to treatment ones. This in turn would result in answers unwanted by the seeker, such as information on measles or the need to avoid TBE, and a consideration of the portal to be "incompetent".

By the "game-theoretic" view of the interaction, we are able to identify several context aspects affecting the possibility of a satisfied information seeker, and the seeker's trust. For instance, when posting the query the seeker may with different probabilities be in certain situations, captured by the search time and seeker location, and expecting specific answers to satisfy the needs expressed by the query and its context. Moreover, the outcome is affected by the ability of the portal to interpret the query given its context, and the seeker need in past similar situations.

At the core of a successful interaction between an information seeker and a portal we find the ability of the seeker to *express herself in the best possible way*, and for the portal to make *accurate interpretations* and *draw conclusions* on the needs of the seeker and the answers to provide. Based on the search logs of the portal and our game-metaphor we are able to create a "model", or *search game*, of the interactions reflecting the seeker's situation when posting the query, the portal's interpretation of the situation as captured by the search log, and the resulting situations of the seeker after obtaining certain answers. However, the seeker situations are not explicitly provided by the search logs, but have to be hypothesised from the given information. If we assume this is possible to achieve, we end up with different situations captured by similar contexts, but where the seeker expects different answers. For instance, as in the case of Julia, searching with symptoms most probably indicates a greater interest in treatments than in preventions. Hence, the model helps us *identify* and *describe* potentially challenging types of interactions, and their impact on the public's trust in portals as a means to obtain advice on health care.

To summarise, it is important for health portals to be able to interpret information seekers' needs as well as possible, and if needed, support the seekers in refining their queries. By a game-theoretic view of the interaction we are able to capture and describe the challenges, and possible solutions, in an accessible way and provide a theoretic framework for analysis and implementation of portal improvements, including trust enhancing portal behaviours.

Today, there is no cure for TBE and immediate and adequate care is important to alleviate the symptoms and avoid severe complications such as paralysis and decreased mental capacities.

1 Introduction

In the report *Svenskarna och internet 2013* 'The Swedes and the internet 2013', it was estimated that 89% of the Swedish population have internet access, 69% of these sometimes search for health information and 4% do it daily (Findahl 2013). From an international perspective, an American study by the Pew Research Institute (Fox 2013) and a study of five European countries (Norway, Denmark, Germany, Greece, Portugal) by Kummervold and Wynn (2012) support these estimates. Of the American health seekers, 77% start their search using a general search engine and 13% access health focused sites as a starting point (Fox and Duggan 2013). Based on these numbers, we may hypothesise that successful Swedish health portals have to be able to support the more than 5% of the Swedish population who sometimes use health-focused portals as their primary internet health information provider, and the up to 40,000 citizens with reoccurring daily visits.

Even though there is a large public interest in health related portals, only a few countries, including Sweden and Denmark (sundhed.dk 2011), have established official national portals to provide citizens with abilities to search for information and manage their health care interactions (CeHis 2012). The aim of the official Swedish health portals[6] 1177.se and vardguiden.se is to promote health and empower the public by reliable and easily accessible health information, and facilitate the health care process and the patient's contacts with care givers (Mannberg 2013). The portals also offer the possibility to compare different care givers and fees (Hyttsten 2012).

[6]By a *portal* we mean any internet based solution providing an interface to browse and search for information on a given theme, e.g. health care. To *search*, or query, is to post one or more words, called a *query*, to the underlying search engine of a portal. The *search engine* will then make use of the knowledge encapsulated by the portal to provide one or more *answers* to the seeker. The search is, more or less, interactive with the search engine, for instance, proposing potential refined queries, based on the posted query terms, to guide the seeker in her task. The *interactions*, i.e. queries and information such as seeker location, used device and number of provided answers and chosen answer, may be stored in so called *search logs*.

The first portal, called 1177.se, is a common portal for Swedish regions and counties, and 1177 is also the official national telephone number for health information and advice. The Stockholm Health Care Guide, which can be accessed as vardguiden.se, is the official health information portal of the County of Stockholm, used mostly by people living in the Stockholm area. Vårdguiden is available on the internet, as a magazine and as a telephone service. In January 2013 the Stockholm County Council reported that vardguiden.se had 2 million visitors per month and 1177.se had 3 million visitors per month (SLL 2013). As of November 2013, the two portals have merged into one called 1177 Vårdguiden, sharing interface and search engine. However, in this work we treat the portals as different entities, since in the past they had slightly different interfaces and search engines, reflected by the herein studied search logs.

In addition to these estimates on the number of portal visits, no detailed statistics on the use of the Swedish portals has, to our knowledge, been established. However, according to American studies, the majority of the topics covered by U.S. health information seekers concern specific *diseases*, *treatments* or health *professionals* (Fox 2013). Among these 35% try to figure out what they, or someone they know, may suffer from, 20% access rankings and reviews to help them choose care providers, and 10% want to read about other people having similar concerns as they have themselves. Moreover, more than 25% of the adults have a smartphone which they use for health related information search, and half of them have apps to track or manage their health.

Hence, there is a trend towards more of the initial contacts between patients and health care to take place via internet portals providing information on symptoms, treatments and care administration. These interactions are carried out using devices such as computers, smartphones and tablets, each with its own technical challenges and way of use. Consequently, it is crucial for portal providers to understand questions such as *how* seekers express themselves, *what* they want to express, and *when* and *why* they express themselves in certain ways. This understanding is even more important for providers of official portals, since as a result of today's and tomorrow's health care becoming increasingly computerised there is a risk of some groups of people being excluded, or that some needs are overlooked, with in the range of 40% of the Swedish population not feeling they are part of the evolving "information society" (Findahl 2013: 59).

1.1 Research questions

The aims of the thesis are to present a *framework* to *describe* the types of *search* and *answer strategies*, inducible from *search logs*, used at two *official Swedish health information portals*,[7] and how it can be used to *guide health information providers* on how this type of portals are to function. Moreover, the framework, called *search games*, and this thesis have a basis in the assumption that a fruitful interaction between seekers and portals depends on the ability of a portal to establish and maintain seekers' *trust* with the *language* as expressed by queries and answers as its facilitator.

The thesis focuses on *natural language processing*, i.e. the use of computers to interpret and generate expressions in human (natural) language, as a means to study seeker–portal interactions. For instance, *how* do users express themselves in searches, *what* do they want to express and *when* and *why* do they express themselves in certain ways.

One of the most fundamental questions when someone intends to search for information on health portals is *how* to express oneself to retrieve the most "useful" information. Should one use a similar type of expressions as when searching with Google, or will the seeker have to adapt to underlying search engine differences? According to Liu et al. (2013), the more familiar you are with a search engine the more complex and question-like queries you tend to post. Thereby, we might expect to see differences in search behaviour between information seekers belonging to the "Google generation" and inexperienced users.[8] Hence, understanding the *stylistics*, i.e. how "authors" express themselves and common linguistic features of different types of writing, used by information seekers in their interaction with health portals may give valuable insights to its providers. For instance, are acronyms commonly used, do seekers post complex queries as in the case of experienced Google users, and do they use terms like *tuberkulos* or *tbc* when searching for information on tuberculosis. Another important aspect is how they have expressed themselves when they do not obtain any answers, and by *error analysis*, i.e. the study of the types of linguistic errors people make when learning a new language, we may gain insights into the type of problems they face, e.g. unfamiliarity with the language of medicine and its terminology.

As important as how users express themselves is the question of *what* they want to express. In the context of health informatics there is a tradition in the

[7] The search logs we have studied are from 1177.se covering the County of Västra Götaland, and from vardguiden.se on the level of counties in Sweden, including a more detailed log of the Stockholm County.

[8] Health information versus other types of internet search has also been studied by us in (Moradi et al. 2014).

spirit of Linnaeus of organising terms into terminologies and hierarchies. For instance, if one searches for information on lung cancer one might also be interested in information on the more general concept cancer, or when searching for information on diabetes one may want to know more about insulin or the pancreas. Hence, being able to map query terms to *semantic*, i.e. defining and describing, concepts may provide valuable query answering information and understanding of searches. The latter aspect is at the core of *infodemiology*, that is, methods to study the "determinants and distribution of health information for public health purposes" (Eysenbach 2006: 247–248). For instance, by studying flu-related searches over time, health authorities can predict spread of the disease and need of care (Hulth 2013). It is important to emphasise that how seekers express themselves is captured by search logs, but what they wish to express has to be induced from this information, and is thereby founded in hypothetical reasoning.

Analysis of health search logs may not only result in an understanding of how and what, but also of *when* people express themselves in certain ways. For instance, searches for information on ticks in the spring could indicate an interest in prevention of diseases spread by ticks, but during late summer reflect an interest in treatments of these diseases.[9] Considering when people express themselves in certain ways includes aspects of both the how- and the what-analyses. Hence, it has to be based on both captured interaction information and hypothetical reasoning.

Finally, understanding these three aspects of searching may help us describe *why* people express themselves in certain ways when searching health portals like 1177.se and vardguiden.se. For example, if one posts symptom terms as a query the interest could be targeted more towards obtaining information on the type of disease one may suffer from than how to avoid it, c.f. 35% of the US population use the internet to diagnose their problems (Fox and Duggan 2013). Similarly, differences in when people search may reveal reasons why. For instance, discussions in media on specific diseases or treatments might lead to changes in query expressions and used terms. This part is obviously the most speculative one, since it depends on all the other aspects.

The aspects of how, what, when and why people express themselves in certain ways when searching for public information at official health portals are facets used in this thesis to, hopefully, provide insights into

- improved health portal *usability*, by increased understanding of how, when and why users express themselves in certain ways at health portals

[9] Change in search behaviour over time has been studied by us in (Eklund 2012a).

- how the way information seekers express themselves may reveal information on their *health status*, hence the type of information that would be most useful to the seeker
- understanding of the relation between queries and answers to establish and maintain seekers' *trust* in health portals

Health related information search is by definition an *interaction*, or act of communication, which takes place by written queries and answers between an information seeker and a provider utilising a search engine. Hence, to study the topics above we would benefit from a *framework* which allows modelling the outcomes, the interaction and the situations enclosing these. It should also facilitate both *descriptive* and *predictive* analysis originating in interaction transcripts as described by search logs. To achieve this, we introduce a *situation-anchored game theory* framework, called *search games*, inspired by the work by Parikh (2010) on mathematical models of communication acts, and the paper by Parfionov and Zapatrin (2011) on a game-theoretic perspective on web search.

1.2 Thesis outline

Following a brief background on e-health (chapter 2), we will in part I (Conceptual framework) introduce a model, called *search game*, of health information search inspired by situation-anchored game theory. This will then be used in part II (Case study) to describe the interactions between information seekers and the two main public Swedish health portals 1177.se and vardguiden.se. The description will address questions on how and in which situations the seekers and portals act in certain ways. This part also includes discussions on how this understanding provides insights into properties of a portal which we believe will increase seekers' trust, and the impact of a potentially changed user behaviour by an increased use of smartphones. In the last part we will summarise our efforts, and address the question if the future is already here with portals able to replace human interactions with care providers and facilitate the health care process.

1.2.1 Conceptual framework

The process of searching for information at a web site, for instance a health portal like 1177.se, can be seen as an *interaction* between two *actors*; the information seeker and the underlying search engine of the portal. The interaction

begins with the seeker making a *move* by posting a query at the portal's search interface. The query is often a result of both the interest and knowledge of the seeker and her understanding of how the search engine "functions". The portal responds by trying to provide, in its "opinion", the best possible answers to its interpretation of the query. This interaction continues until the seeker either considers her query answered or views the portal unable to provide the needed information. The sequence of seeker moves can be seen as an unfolding of her *search strategy*, and similarly the moves of the portal as the *answer strategy*. By this, the aim of both actors is to choose strategies which satisfy the information needs of the seeker. To a health information provider, this implies choosing a strategy which best matches the needs and knowledge encoded in the chosen search strategy.

In part I (Conceptual framework), we introduce a *game-theoretic* view, incorporating the enclosing situations, of searches along the line of reasoning described above where the seeker and the portal interact by making moves following strategies to achieve the aim of the best possible outcome for both parties. Since this view of information search may not be mainstream,[10] we begin by providing an introduction to game theory (chapter 4), originating in the early twentieth century interests in parlour games, economics and formal descriptions of science, leading to our theoretical framework used as a basis for reasoning in the rest of the thesis. As we will show, a game-theoretic view of search provides a simple relation between search logs and the notion of strategies in a game, where the former play the analogous role of chess transcripts for a game of chess. Moreover, by analysing these used strategies we show how to infer preferences for different sequences of moves. For instance, if you enter the search terms *fever cough*, you may receive answers related to influenza and that may be what you were interested in, or in terms of game theory, an *equilibrium,* i.e. a state of affairs where none of the involved parties would benefit from moving to another state, has been reached.

Query terms like *fever* and *cough* are both examples of symptoms related to diseases, and generally one may expect to obtain information on the underlying disease or its treatment. Hence, the portal should "interpret" the query terms taking into consideration the most plausible *context*, e.g. that symptom words indicate an interest in treatment answers. Hence, being able to map the syntactic expressions to semantic concepts, or their meanings, can provide insights into patterns of searches, and expected types of answers. The semantic analysis can then be used to study the scenarios where certain search patterns

[10]Game theory has also been used to study pragmatics by, for instance, Parikh (2010), especially communication acts. However, we will not elaborate on this research since, even though search may be seen as a type of communication act, Parikh's framework is far more detailed and formal than needed in our work.

are more common than others. For instance, searches where pairs of symptoms and diseases are found in the search log might be indicative of treatment scenarios, and we introduce the notions of *search scenarios* and their relation to search games, cf different types of chess openings and endings.

Then we briefly discuss how search logs can be used to induce *payoff* functions describing the *preferences* among (types of) answers for given (types of) queries. Thereby we have established a conceptual framework, based on available search logs, allowing both descriptive and predictive analysis of health searches. We end the first part of the thesis outlining how it is used in a case study to address questions originating in the need for a portal to maintain seekers' trust both in cases of queries without answers and ones with many answers. As a consequence, we are able to introduce a number of principles of a trustworthy portal, based on the behaviours of the seeker and portal as reflected by their querying and answering.

1.2.2 Case study

In part II (Case study) of the thesis we use the introduced framework to establish a description of how the official Swedish health portals 1177.se and vardguiden.se are used, aiming at an understanding of how aspects like seeker demographics, degree of mobility and time of search affect the way information seekers express themselves in the hunt for health advice and care. We also study how the interactions between the seeker and portal may indicate patterns of behaviour of interest for improved portal support, especially the potential impact on seekers' trust in a portal's ability to provide adequate information and advice. The analysis addresses two important settings of interaction:

- When a query results in *no answers*, and
- When a query results in *too many answers*, according to the seeker

These aspects are important, since in the theoretically best of worlds a health portal should be able to interpret a seeker's query so well that it only has to provide one single answer exactly addressing the seeker's need. Obviously, this is not the case today, and might never be, but by the increased mobile use of information portals and political interest in using internet as the first point of contact between potential care seekers and providers, it becomes increasingly important for solutions such as portals to provide the right type and amount of information at the right point in time.

Our analysis addresses topics such as the impact of context on the number of answers and characterisation of common reasons for seekers ending up with no answers (chapter 9), and some of our reflections are:

- There is a thin line between a portal utilising search location and time information to provide better seeker support, and ending up with a seeker considering the portal not to be trustworthy due to not receiving any answers since the constraints may have been too restrictive when location and time are included.
- Many of the queries without answers result from seekers not knowing the "language" of medicine or the portal.
- The answers to a query tend to change over time in a way increasing the risk of seekers considering the portal's behaviour to be "irrational".

In the case of queries ending up with too many answers, we focus in chapter 10 on characterising the ones resulting in seekers having to, in theory, browse more than five answer gists before deciding on an interesting one. In this case, some of our reflections are:

- In general, there is not a clear relation between the expected and the actual number of answers, possibly resulting in seekers questioning the portal's "competence" and ability to support the seekers' needs.
- There is often a detectable relation between the types of queries posted and answers of interest, but semantic annotation do not provide as much support to identify these as possibly expected.

When combining the findings of these topics in chapters 9 and 10, we are able to present a number of principles to be adhered to by a trustworthy, in the eyes of an information seeker, (health) portal (chapter 11).

1.2.3 Summary and conclusions

In the last part of the thesis we address the question if portals can replace human interactions with care providers and facilitate the health care process, and conclude that much work remains before this will be the case, especially on the use of semantic annotations and contextual constraints. We also raise, as a consequence of our game-theoretic perspective, the questions if and how a rational portal may be realised and its role as an actor able to not only react to seekers' behaviours, but also by its answering strategies to influence and change the search strategies, i.e. the seekers' behaviours.

1.3 Contributions

During the last decade, the first point of contact between a potential care seeker and health care has shifted from a physical meeting to an internet interaction, often via a smartphone or other mobile devices. The seeker can often be in distress, trying to determine, by searching the internet, if it is necessary to seek care. Many have no, or very limited, medical knowledge, and need support from the health portal. The queries are often ambiguous and do not always reflect differences in the seekers' information needs. For instance, if the seeker types the query *stroke* she may be interested in treatments, how to prevent the disease or want to read about other people's experiences of the disease. Hence, the health portal is confronted by information seekers in different circumstances, often expressing themselves in unclear ways, and with different expectations. Consequently, the increased use of mobile devices and situations with seekers in distress raise the question on how portal providers may be able both to better *model*, or describe, the user behaviour and to *predict* the impact of changes in search algorithms.

Inspired by the work by Parfionov and Zapatrin (2011) on a model of information retrieval, and Parikh (2010) on semantics and pragmatics, both with a basis in game theory, we show how these outlined challenges can be addressed, and introduce a framework for descriptive and predictive analysis in the setting of health search. By the foundations of the framework we are also able to address questions regarding preferences of answers and the trustworthy (health) information portals.

By the analysis of the use of two Swedish health information portals, and a re-use of research in game theory, error analysis and stylistics, we hope to have been able to provide

- A *model* for descriptive and predictive analysis of (health) *information search*, allowing, in theory, both automatic induction of preference relations and what-if analysis of changed behaviour of seekers and portals
- A *definition* of seeker and portal strategy *preference* facilitating studies of *trust* in (health) information search based on search log data
- An in-depth *analysis* and description of the search carried out at the official Swedish health information portals 1177.se and vardguiden.se, with emphasis on *queries without answers* and *queries with many answers*
- An introduction to error analysis and stylistics for information search analysis

- Examples and discussion on the use of annotation sources as the UMLS for information search analysis and improvements
- Discussion on properties of search logs to facilitate information search analysis
- A set of *principles* a *trustworthy* (health) *information portal*, in our opinion, should adhere to

The content of this thesis, or related topics, has been partially presented in the following papers:[11]

Oelke, Daniela, Ann-Marie Eklund, Svetoslav Marinov and Dimitrios Kokkinakis 2012. Visual analytics and the language of web query logs – a terminology perspective. *The 15th EURALEX International Congress (European Association of Lexicography)*, 541–548.

Eklund, Ann-Marie 2012. Tracking changes in search behaviour at a health web site. John Mantas, Stig Kjaer Andersen, Maria Cristina Mazzoleni, Bernd Blobel, Silvana Quaglini and Anne Moen (eds), *Quality of life through quality of information Proceedings of the 24th European Medical Informatics Conference*, Volume 180, 858–862. IOS Press.

Eklund, Ann-Marie and Dimitrios Kokkinakis 2012. Drug interests revealed by a public health portal. *Proceedings of the SLTC-Workshop: Exploratory Query-log Analysis*.

Eklund, Ann-Marie 2012. Are prepositions and conjunctions necessary in health web searches? *Proceedings of SLTC 2012 The Fourth Swedish Language Technology Conference*, 23–24.

Eklund, Ann-Marie 2012. Why query annotations may help in providing accurate public health information. *ESAIR'12: Proceedings of the fifth workshop on Exploiting Semantic Annotations in Information Retrieval*, 5–6.

Kokkinakis, Dimitrios and Ann-Marie Eklund 2013. Query logs as a corpus. Andrew Hardie and Robbie Love (eds), *Proceedings of the Corpus Linguistics 2013*, 329–330.

[11] Some of the results in the papers may differ slightly from the ones presented in this thesis due to improved analysis methods or different data sources.

Eklund, Ann-Marie 2013. Mobility and health information searches – a Swedish perspective. Christoph Ulrich Lehmann, Elske Ammenwerth and Christian Nohr (eds), *Proceedings of the 14th World Congress on Medical and Health Informatics*, Volume 192, 1079–1079.

Eklund, Ann-Marie 2013. On challenges with mobile e-health – lessons from a game-theoretic perspective. Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, Rajeev Rastogi and Fabrizio Silvestri (eds), *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM'13*, 1249–1252.

Moradi, Farnaz, Ann-Marie Eklund, Dimitrios Kokkinakis, Tomas Olovsson and Philippas Tsigas 2014. A graph-based analysis of medical queries of a Swedish health care portal. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 2–10.

The research presented in this thesis, and related papers, has been financially supported by the *Graduate school of language technology* (GSLT), *Centre for language technology* (CLT), *Wilhelm och Martina Lundgrens Vetenskapsfond 1* and *Filosofiska fakulteternas gemensamma donationsnämnd*.

If not stated otherwise the results presented in this thesis are the work of the author.

2 From traditional health care to e-health

By the birth of the internet in the early 1990s, the life of a large part of the world's population has changed with people "living" part of their lives on the "net", carrying out everything from shopping to dating. The internet also offers the possibility to search for information on any topic in a way never seen before.

Obviously health care, being part of daily life, has also become part of the "net". For instance, in Sweden several official projects under the umbrella of *Nationell eHälsa – strategin för tillgänglig och säker information inom vård och omsorg* 'National eHealth – the strategy for accessible and secure information in health and social care' (Socialdepartementet 2010) have been initiated to make public health care accessible via the internet to the Swedish population.

According to Josefsson (2011: 22), in parallel with the "democratisation" of health care during the 1970s and 1980s, a discussion evolved on the individual's role in health care. This led to the belief that care seekers should be able to choose their care givers, but also to discussions on the quality of care and communication between care givers and takers. Over the last few decades, the role of the patients has changed from passive care takers to active participants who take responsibility for their own health care – the patient has become "empowered". She has become a "collaborator" involved in the care, with both will and ability to critically review, compare and choose care givers and treatments.

To take an active part in their own care, patients need to be informed and many become, more or less, "experts" on their own disease. This has in a large part been made possible by the internet as a source for learning about, for instance, a disease and its treatments, with patients searching for facts on diseases, treatments and the health care system. The internet offers both official and commercial health portals, allowing information seekers to learn more about diseases and treatments as well as managing their own care. It also offers portals for professionals on almost any possible aspect of medicine and health care. Some of these, like the Swedish medication registry FASS (LIF 2013), al-

low both professionals and the general public to search for information on the effects and side effects of medications. On the internet people may also read other patients' stories and communicate with people with similar experiences to receive and provide information and support.

2.1 The changing role of information seekers

When patients have gone from being passive consumers to active participants in their own health care, their need to learn about and understand their situation, for instance if they need to seek care, available treatments and side effects of medications, has turned them into web "crawlers" searching for relevant information. Thereby, active participation in ones health care also impacts the way sources are used to gain information.

Information search can be *active* or *passive* (Josefsson 2011). Finding information by chance when searching for other things or not searching at all is called passive search. Active search can be actively searching for new information, or returning to familiar sources to see if there have been any updates and new information has been added ("ongoing" search). Active search often starts with a search query at a general search engine, resulting in a large number of hits for the information seeker to go through. For instance, an American study estimates that less than one out of five health related information searches start at a dedicated site (Fox and Duggan 2013). In this context it is also interesting to note that even though the interest in health information has increased over the last years in Sweden, the number of people searching for this type of information on a daily basis has decreased (Findahl 2010, 2011, 2012, 2013), figure 1, possibly indicating that an active search has been replaced by a passive search with many other forums and news portals containing more and more health related information.

Information seekers often find it difficult to determine if information found online is useful and accurate, and most information seekers do not check if the source is reliable (Josefsson 2011). Many seekers prefer portals recommended by someone they trust, e.g. a medical professional. The appearance of a portal is important and a simpler looking one makes the seekers more sceptical. Another common way to determine what seems reasonable is to look through many portals and compare their contents. However, the most important factor is the origin of a portal, and it is considered trustworthy if it was produced by e.g. Socialstyrelsen (The National Board of Health and Welfare), a hospital or a university.

These findings are interesting in the light of a minor part of health related searches to use this type of portals as their primary starting point, and a his-

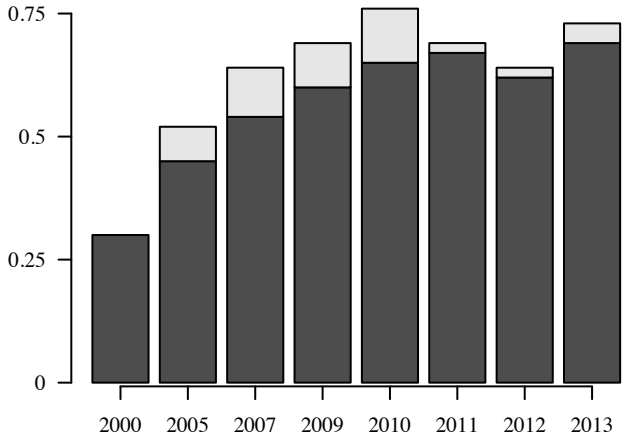


Figure 1: Proportion of Swedish internet users who sometimes (black) and on a daily basis (grey) search for health information.

torically active search behaviour to possibly be replaced by a more passive one via general news portals and social media. The role of the internet as a health information provider is also of interest noting that, according to Fox and Duggan (2013), less than 40% of the people using the internet for "diagnosis" visited a professional care giver to confirm their findings. However, it is important to stress that people have always been self-diagnosing and -treating health problems, but today a third of the population have added the internet to their "diagnostic toolbox".

To summarise, the internet has become an important player in the health care system, and official portals and forums have to ensure that this does not lead to misinformation and, in the worst case, inadequate health care.

2.2 Challenges for health care

As the population gets older and chronic diseases become more common, the health care needs increase (CeHis 2012). The possibility to compare and choose care givers also means higher expectations on the service, quality and availability of health care, but at the same time there will most likely be no increased funding to meet the demands. Eriksson and Majanen (2012) predict that patients – and personnel – will be more mobile and willing to seek care in other countries, which leads to international competition for care givers and further demands on information providers.

As discussed above, more and more people use the internet for health information, and care givers need to have the resources to meet the informed public – or in some cases misinformed (Josefsson 2011), since people sometimes misunderstand the information they find online, or it is not applicable to their specific case. This can cause problems in patients' encounters with the health care system and discussing information found online may take time from other matters during the consultation. To facilitate this discussion with care seekers, it is important that care givers know what information can be found on the internet and what is discussed online.

Another challenge for the health care system is that not everyone can, or wants to, use the internet (Josefsson 2011). Even though Sweden is one of the countries in the world where most people have access to computers and internet, there are still those who do not have the technology, knowledge or skills necessary for efficient internet use. For health search, not only technical knowledge is necessary, but language skills are also important, both English and the medical language, and knowledge of the health care system can be useful. This difference in resources to use the internet, the so called *digital divide* (Josefsson 2011: 130), is a problem which grows larger with the increasing demands on the public to be more and more active in their own health care, and it can cause inequalities in health care, where the more informed people have better possibilities to receive, or question, care and treatments – the digital divide causes a "medical divide".

For those who can and want to adopt the new technology Eriksson and Majanen (2012: 133) predict good opportunities for individuals to shape their own health care, with respect to both prevention and treatment – "det sjukhusfria samhället" 'the hospital free society', where the planning is central but the execution is offered by many different means and actors. According to the authors, the archetypes of *explorers* and *avantgardists* will make up almost 50% of the population in the year 2035. These groups are especially interested in a healthy lifestyle and efficient care, and willing to try new forms of disease prevention and treatment. Hence, in twenty years, self-care, self-improvement and preventive health care will be the leading trends, with internet as a "vårdcoach" 'care coach'.

To summarise, the health care system is in a state of change resulting from a population more active both in their choice of care givers and using internet as a source for information on diagnosis as well as treatments. Even though this "freedom" may be beneficial to some groups in society, it also leads to a risk of a "medical divide" between those groups and the up to half of the population which is not able or willing to follow the trend towards the "hospital free society".

2.3 The internet and health information

With a society moving towards becoming "hospital free" with citizens using the internet for tasks reaching from management of their health care appointments and prescriptions to self-diagnosis and treatment, understanding the expectations and use of the internet is crucial.

By the report *Svenskarna och internet 2013* 'The Swedes and the internet 2013' (Findahl 2013), it is clear that not only do Swedes use internet on a regular basis but that it is also often used for searching for health related information, table 2.1 and figure 1. In Sweden, according to Rahmqvist and Bara (2007), people's health information seeking on the internet tripled between 2000 and 2005 if consideration is taken to the general increase in internet access and use during this time period. The authors also found that women searched more than men, and it was mainly the young and middle aged who used the internet to find health information. In the 2010 study of Swedish internet use (Findahl 2010) this was still the case, and it was also noted that well-educated are more keen health information seekers than poorly educated. However, when the study was adjusted for internet use in general only the education aspect was noticeable. Moreover, the search for health related information seemed to be higher among internet users with health problems.

Swedes with internet access	89%
People aged 12 or over who have used the internet daily	74%
Internet users who sometimes search for health information	69%
Internet users who search daily for health information	4%
Non-users who sometimes ask others to search for them	55%

Table 2.1: The Swedes and the internet 2013.

That internet use is increasing has been supported by, for instance, Wangberg et al. (2009), who found an increase in internet health seeking in Norway from 2000 to 2007. That the importance of the internet as a source for health information is a global phenomenon has also been noted by a study of five European countries (Norway, Denmark, Germany, Greece, Portugal) by Kummervold and Wynn (2012). However, this study shows a higher use in northern Europe than in the south.

As discussed above, the intention of the "hospital free" society is both better quality and use of the health care system, and this has been supported by Wangberg et al. (2009), who conclude that there is a potential for using the internet for health promoting purposes. However, Weaver et al. (2009) raised the question of how effective an asset it actually is for health promotion and

disease prevention, considering the differences in health information seeking behaviour in different population groups. For instance, certain groups may seek their information on sites like blogs, wikis and social forums as part of their "normal" internet use, thereby being at risk of basing their health related decisions on unverified and possibly incorrect information.

Moreover, Weaver et al. (2009) showed internet medical information seekers to report higher health care system use. This view has also been supported by Palen et al. (2012), who show that online access to clinicians and medical records results in increased use of health care services, both visits and the use of telephone service. However, in a study by Medlock et al. (2013) on elderly internet users in the Netherlands, 85% of the respondents had used the internet for health information. This had caused 51% to make life style changes, but fewer of them had acted on the information in other ways, e.g. deciding to seek care (34%) or discussing the information they had found with their doctor (30%). The study by Fox and Duggan (2013) found that among the 35% *online diagnosers*, i.e. those who use the internet to diagnose what condition they or someone else may have, 53% talked with a clinician about their problems and 41% had their diagnosis confirmed.

Not only is it unclear if and how the internet affects the use of health care, but also in the absence of a diagnosis people search for information to confirm their initial hypothesis – and often manage to find "confirmation" (Keselman, Browne and Kaufman 2008). According to the authors, lack of domain knowledge is a problem, leading to difficulty locating, understanding and applying health information. Weaver et al. (2010) found an association between health self-assessment and the type of health information searched for (wellness, illness etc). Illness-information seekers have more negative health indicators and health risks, such as health status, quality of life, BMI, smoking, physical activity, etc. Wellness-information seekers, on the other hand, have more positive health indicators and less health risks.

Another dimension of health related information search is if it takes place on behalf of someone else. Sadasivam et al. (2012) found differences between seekers depending on if the information sought was for themselves, so called *self seekers*, or if they were searching on someone else's behalf, called *surrogate seekers*. Of all seekers, 56% were surrogate, which is slightly greater than the less than 50% surrogate seekers identified by a Pew Research Institute study (Fox and Duggan 2013). The typical surrogate seeker, according to Sadasivam et al. (2012), is married, a parent, in good health and often an informal care giver. Surrogate seekers are most likely searching for specific diseases or treatments, medical facilities or experimental treatments. Not found among the surrogate seekers are what we might call the *social seekers* (Reis, Church and Oliver 2012), those considering health information search to be

a social activity, especially today when any information source is reachable from a smartphone or a tablet. For instance, in Sweden 59% of the population aged 16–74 access the internet via mobile devices (Digitaliseringskommisionen 2013) and mobile devices such as smartphones have become more important when searching for health related information (Church and Oliver 2011).[12] By their smaller size, the interaction between users and portal has to be different in comparison to when the seeker uses a computer. The portals must be better at predicting the needs of the user to provide adequate information without forcing the user to browse long indexes of suggested pages of information. For instance, searching the Swedish health portal 1177.se during the period June 2010–September 2011 with the query *stroke* results in an average of 204 suggested answers, but in a search log of the portal only 21 were chosen, with a rank[13] reaching from 1 to 15 (average 3.7). Hence, the increased mobility not only impacts the social setting where searches take place, the expected interaction with the information portals, but possibly also the type of information searched for.

The majority of the topics covered by health information seekers in the U.S. concern specific diseases, treatments or health professionals (Fox 2013). 35% try to figure out what they, or someone they know, may suffer from, 20% access rankings and reviews to help them choose care providers, and 10% want to read about other people having similar concerns as they have themselves. Moreover, more than 10% of the adults have smartphone apps to track or manage their health. Since more than 50% of the seekers are interested in information to help them make decisions based on what they may suffer from and available care providers, the ability to *trust* the information and its providers becomes important. According to the presentation "eTjänster inom hälsa, vård och omsorg" by the consultancy firm Kairos Future (2012), in 2011 88% of the Swedes had at least a rather high trust in information from sources like 1177.se, in comparison to 94% with a similar trust in information obtained from a doctor. However, only 61% had a rather high trust in health information they found on the internet. This number was higher among people aged 16-25 where 26% would follow advice given by "computers" in comparison to 11% of the population in general. Hence, even though official information providers are more trusted than internet in general, it is still lower than in comparison to a doctor's visit.

To summarise, the increased health information seeking makes the internet a potential means for health promoting and disease prevention, but it also seems that more searching can lead to greater use of the health care system,

[12]Mobility and health information search has also been studied by us in (Eklund 2013a).

[13]The *rank* of a suggested answer is its position in the list of suggestions.

with both visits and the use of telephone services increasing. Information seeking is often a social activity, and seekers often start out with a hypothesis which they want to confirm, and often – even when the hypothesis is incorrect – they are successful in finding information that they can interpret as a confirmation, thereby possibly just enforcing their own view of their health status.

2.4 The health portals 1177.se and vardguiden.se

The internet sites 1177.se and vardguiden.se are two official Swedish health portals, providing information on e.g. diseases, treatments, health care facilities and patients' rights. The information is written by medical professionals and is available in different languages. The portals also provide help to find local health care providers, and through the portals it is possible to contact personal care givers to make appointments, renew prescriptions, etc over the internet using the service *Mina vårdkontakter* 'My care contacts'.

The portal 1177.se started in 1998 and has successively been developed and new contents have been added (Hyttsten 2012). In December 2010, it became the common national portal for health and dental care, containing both general and regional information. The number 1177 is also the official national telephone number for health information and advice. Each regional part of 1177.se provides local information on, for instance, care givers or events. The portal contains text, illustrations, animations, films for children and an anonymous querying service where doctors and nurses answer questions.

The aim of 1177.se is to promote health and empower the patient by providing reliable and easily accessible health information, and facilitate the patient's contacts and dialogue with the health care system (Mannberg 2013). Some of the goals stated for 2013 by the Center för eHälsa i samverkan CeHis (Centre for eHealth in Sweden) are to have 4 million visitors/month, that 35% of the population know about the web site, and that 30% use it regularly for e.g. finding contact information for care givers (CeHis 2012). Another goal is to improve the querying service and answer 10,000 questions a year.

The Stockholm Health Care Guide, vardguiden.se, is the official health information portal of the County of Stockholm with nearly 2.1 million inhabitants. It is used mostly by people living in the Stockholm area, but also by people from other parts of Sweden. Vårdguiden is available on the internet, as a magazine and as a telephone service. A mobile version of vardguiden.se started during 2010 (SLL 2010).

In January 2013 the responsibility for both information services was transferred to the Stockholm County Council. The new e-health service started in November 2013 under the name 1177 Vårdguiden. In this work we treat the

portals as different entities, since they in the past had slightly different interfaces and search engines, reflected by our search logs. However, the user interface has, as we view it, not changed substantially, especially considering the search related facets of interest in this thesis.

When an information seeker accesses the portal 1177 Vårdguiden, she finds a query form allowing her to submit queries, figure 2.[14] The front page also contains shortcuts to areas covering *Fakta och råd* 'Facts and advice', *Hitta vård* 'Find care' and *E-tjänster* 'E-services' (make an appointment, cancel or reschedule an appointment, renew prescriptions, view medications). The portal suggests queries while the seeker is typing, for instance *stelhet i nacke* 'stiffness in neck' as an alternative to *stel i nacken* 'stiff in the neck'. The portal uses cookies[15] and internet protocol information to choose a default county/region to narrow down the answers.

The results provided by the portal are focused on diseases, treatments and care givers, which is clear by the division of answers into groups, figure 3, with the tabs *Alla träffar* 'All hits', *Artiklar* 'Articles', *Mottagningar* 'Care givers', *Frågor och svar* 'Questions and answers', *Nyheter* 'News' and the number of results for each tab. The answers are most often found in one of the categories 'Articles', i.e. facts on diseases or treatments, or 'Care givers', which thereby make up the bulk of the answer set of 'All hits'.

If we click on the second proposed answer we are directed to the "fact sheet" on *hjärnhinneinflammation* 'meningitis', figure 4. If we instead would have submitted the query *vaccination* and chosen the result set corresponding to health care units (*Mottagningar*), we would have been presented with a list of care providers in the given region of Sweden, figure 5.

[14] The pictures present the current merged 1177 Vårdguiden interface, but this is, more or less, the same as before the merging of the portals.

[15] A *cookie* is a piece of information automatically stored on a user's device to allow application re-use in later sessions.

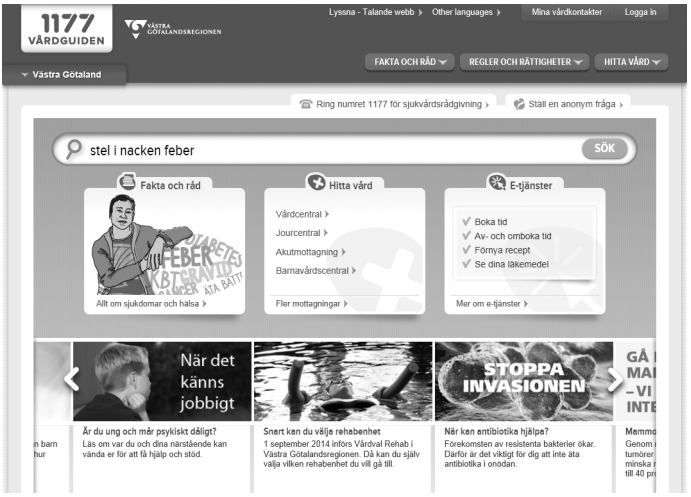


Figure 2: Main interface of 1177 Vårdguiden.

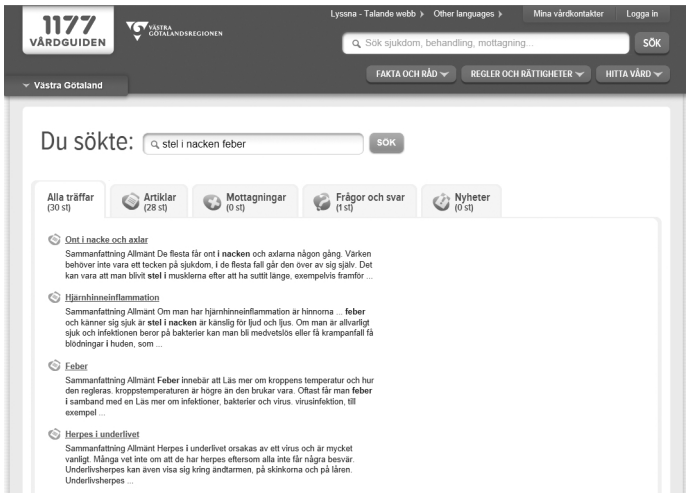


Figure 3: Result page at 1177 Vårdguiden.

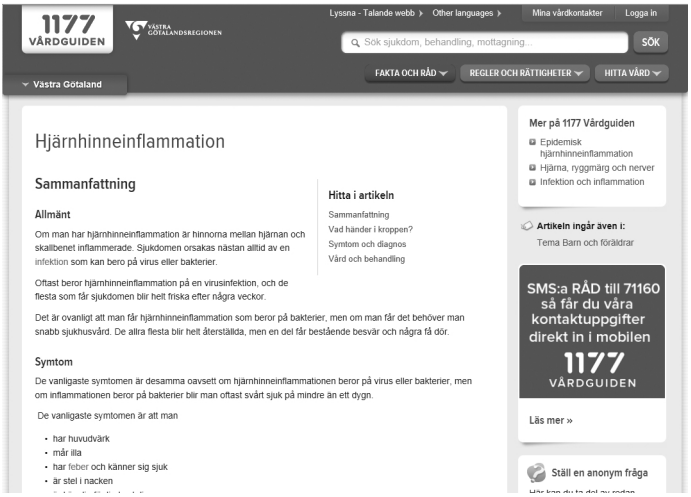


Figure 4: Fact sheet at 1177 Vårdguiden.

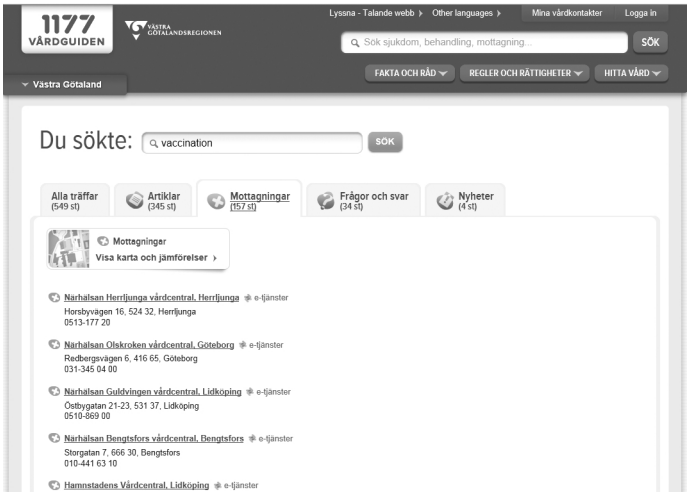


Figure 5: Health care units at 1177 Vårdguiden.

Part I

Conceptual framework

3 INTRODUCTION

In this part of the thesis we begin by elaborating on the benefits of establishing a *conceptual framework* as a basis for our research (section 3.1). However, choosing a framework is not a trivial task and, as concluded in section 3.3, many different approaches have been taken to describe interactions between information seekers and portals. In the preparation of this thesis we were, unintentionally, influenced by the way we describe our efforts to laymen, with discussions along the lines of "first the seeker..., then the portal tries to figure out what the seeker... and it will... If the seeker is happy, then... Otherwise...". Hence, it resembles an unfolding of a sequence of "moves", as in a game of chess, where the "players" try to figure out the other participants' intentions by their moves. Since this was a way not only to describe our research to others, but also to organise our own thinking, we decided to pursue this metaphor of information search and benefit from others' approaches using similar metaphors in their research. In essence, a framework should be able to provide a common way to *describe* different aspects, allow *re-use* of already established concepts and properties, allow *abstraction* from specific instances to patterns and theories, and provide a basis for *predictive reasoning*, at the same time as treating seekers and portals as equally important and active actors in an interaction resembling a play of a *search game*.

As presented in section 3.2, the concepts of *query*, *meaning* and *situation* will all play important roles in our framework, and based on a philosophical view of language and its role in the context of communication we try to motivate not only the importance of the actual query in a fruitful information search, but also its meaning in relation to the needs and expectations of the seeker and the situation wherein it takes place. As an introduction to the theoretical foundation of the framework, we end this chapter with section 3.4 briefly introducing how a search scenario may be described in the forthcoming framework.

Chapter 4 presents the foundations of *game theory* and *situation theory* used as a basis for the framework of *search games* (chapter 5). In chapter 6,

we describe how *search logs* can be turned into search and answer *strategies* and finally into *scenarios* in the search games. The latter may be seen as abstract patterns of behaviour of seekers and portals, cf Sicilian Defence openings in chess or even so called *use cases* in software and interface development (Järvelin et al. 2013), aimed at highlighting important common and differentiating features of the realities described by the models.

3.1 Why a conceptual framework?

One of the more intriguing aspects, in our opinion, of a research thesis is the notion of a *conceptual framework* as a "researcher's map of the territory being investigated" (Miles, Huberman and Saldana 2014: 20) to accommodate the *boundaries*, or purpose, *evolution*, or flexibility, and *coherence* of the research. In other words, provide a framework in the word's literal sense for a research topic.

As discussed by Leshem and Trafford (2007: 96), a framework aims at providing a theoretic overview of the intended research and order within the process, but also to share some of Kuhn's view of a paradigm to convey the way the world is seen through our "perceptions, understandings and interpretations". Moreover, it allows empirical data to be abstracted into patterns and generalisations in the setting of the framework, thereby function as a means to lift the results away from the technical details to a level which "leads to the elucidation of further research questions and implications for additional studies" (Rudestam and Newton 2007: 7). The author Robson (2011: 67) very well describes the reasons for us to use a framework as an important part of the presentation of our research:

> "Developing a conceptual framework forces you to be explicit about what you think you are doing. It also helps you to be selective; to decide which are the important features; which relationships are likely to be of importance or meaning; and hence, what data you are going to collect and analyse."

If the aims to use a conceptual framework are clear, the next challenge is to decide which one to use. As will be presented in section 3.3 there are many ways to describe information search, reaching from cognitive models to, more or less, black-box ones depending on the purpose of the description. Early on in our research we decided to put the information seeker in focus, but not as it is done in many cognitive descriptions at the expense of the portal to be treated as a passive reactive part. Hence, we wanted to emphasise that the portal is as important an actor as the seeker, and that we are dealing with an interaction be-

tween actors. But not one where we are able to drill into their "brains", but are forced to try to understand their behaviours, as well as possible, based on the transcripts of their interactions captured by the so called search logs of queries and answers. With this as a basis, we studied some of the existing models, but were especially attracted by the work by Parikh (2010) on semantics and by Parfionov and Zapatrin (2011) on information search as a game, both using the theory of games as a foundation. When trying to phrase our research interests, but even more important, to be able to describe the interactions between the information seeker and portal in a conceptual way to non-linguists or computer scientists, we found the game metaphor to be very useful. Topics like unanswered queries became incomplete plays in a game, and notions like rational search was expressible in terms of preference relations among strategies for searching and answering, and in all these discussions the seeker and portal were equally important actors with their own interests, wills and strategies to fulfil them.

In our opinion, the chosen game-theoretic framework has been very useful for us to order the process of thinking and defining and describing properties of health information search, but still the framework has not been used at its full potential to function as a framework to study notions like different types of strategies, and their implications, but also concepts like rational behaviour, and how different actions imply other actors' view on rationality, and in the end *trust* among actors. Hence, much work remains to be done with our theoretic framework, called *search games*, as a basis, and this thesis just opens the box for future scientists to use.

3.2 Philosophical view on information search

Influenza is a common disease with symptoms like fever, coughing, sore throat and muscle pain, and the following is a common search scenario in Sweden in February.

> Ann sits by her computer and types the query "influenza" to a health information web site. The portal contains different kinds of information, and answers the query by, among other alternatives, one related to "vaccination" and another one on "how to treat influenza".

Obviously, the portal needs to decide the *meaning* of, or interpret, the *query* to be able to provide the "best" answer. Hence, in a conceptual framework we would like to be able to describe how this interpretation evolves.

A possible path to achieve this could be to follow the tradition of the *ideal language* philosophy. This view of language and meaning, endorsed by, among others, Frege, Russell and Whitehead (IEP 2012), has as its core the idea of *reference*, that a language is *about* the world and the meaning of a word is what it points to in the world. In our example, the word "influenza" would point to the disease influenza, hence its linguistic representation and *conventional meaning*, and we may call this the (formal) *semantics* of Ann's query.

If Ann does not have the flu, she might be interested in general information about the disease or want to know how to avoid it. If she already has the disease, she will not be particularly interested in the information that her local care giver offers drop-in vaccination. She may, on the other hand, be very interested in how to treat the disease, or at least alleviate the symptoms. Hence, Ann's *situation* obviously impacts her preferred answers, but is most probably unknown to the portal, and it is not captured by the ideal language approach. In addition to the challenge of ideal language philosophy to deal with meaning dependent on situation, it also suffers from problems to handle aspects like desire and intention of an information seeker and the ambiguity of her used language. Hence, our framework should be able to take into consideration the *query situation* in addition to the actual queries and answers when describing interactions between information seekers and health portals.

Another view of language, trying to consider these aspects, is the *ordinary language* philosophy, where focus is on *use*, or the activity of communicating, as put forward by, for instance, Wittgenstein, Austin and Grice (IEP 2012). If semantics "underspecifies" the meaning, this philosophy tries to fill the gaps by considering contributions of the ambient circumstances. This is also called the *pragmatics* of language.

If search is viewed as an interaction between actors, that is, a *communication act* between an information seeker and a portal, the aims of the interaction may be said to convey seeker needs and knowledge and for the portal to provide relevant answers. For instance, if Ann in our example posts the query "fever cough" she is most likely more interested in information on influenza than on influenza vaccination, and hopefully the portal signals that it understands these needs by presenting information on the disease. Moreover, when Ann posts a query to the portal, trying to express her information need in the best possible way, it not only depends on the situation at hand, but also on her general background *knowledge*, the portal *interface* and *experiences* from her own previous searches and resulting answers. When she is provided with answers she interprets these in the same context as when posting the query. Hence, why the portal replied with the given answers may not be clear to her. For instance, in the example above, information on vaccination may seem irrelevant when she already suffers from symptoms of influenza. Analogously,

the portal offers the best possible answer to the seeker's query based on its interpretation of the query and its "beliefs" about the seeker. The portal's view of the seeker may be based on its knowledge base (e.g. medical terminologies), seeker demographics and previous searches and answers. Hence, viewing the portal as a "reader" of the query provided by a "writer", it may interpret the query taking into account information about both the writer (demographics) and the reader itself (the knowledge base) and the query. Thereby, the pattern of interaction between the information seeker and the portal is highly related to their ability to interpret the other part's actions expressed in terms of queries and answers, but also implicit aspects related to when the interaction took place and the "style" of expression used by the seeker. Hence, studies of information search will also have to consider the *discourse*, i.e. the explicitly as well as the implicitly considered aspects, of the interactions (de Saussure 2007).

Another perspective on the interaction of a seeker and portal is to adopt Jakobson's (1960) model, see also (Goatly 2008: chapter 1), of the *functions* of language, where an act of communication involves an *addresser* who uses a *contact*, i.e. a channel and medium, to communicate a *message* to an *addressee*. This communication takes place according to a *code* of communication and within a given *context*, i.e. a setting and topic. Jakobson's interest in communication originated in linguistics and the role of poetics. However, we may adapt his model of communication considering information search to describe the different roles the elements of the model play, especially that several of these are important in our analysis. Obviously, the addresser is important and by her way of expressing herself we are able to learn about the "attitude toward what [s]he is speaking about" (Jakobson 1960: 354). For instance, if she posts queries referring to symptoms we may hypothesise that they express concerns about what she is suffering from and how to treat the problem. Thereby, we may say that when focusing on the addresser we are interested in the *expressive* function of the interaction between the seeker and portal. If we instead turn our interest to the portal, i.e. the addressee, it is the *conative* function which is in focus, i.e. the expectations of the seeker in the perspective of the portal. For instance, by posting a query on symptoms, the seeker expects answers on their source and treatment. At the core of the communication we have the message, and the way the code is used to choose terms and combine them in queries makes up the *poetic* function of the interaction, for instance, the use of medical terms in comparison to layman language. However, as discussed the interaction takes place in a context, and its *referential* function is important and impacts both the seeker's expectations and portal's interpretations. Finally, we have also learned how the portal interface, or the *phatic* function, impacts both the querying and the provided answers and that there is a code of communication, or a *metalinguistic* function, which dictates the way of carrying

out information searches. Clearly, all of these six functions of the different elements in Jakobson's model of communication also play, more or less, important roles in a fruitful information search, thereby they are also of interest in a characterisation of the process. Moreover, these functions are also the ones which define and describe any potential stylistics of information search where the information seekers may be viewed as a collective of "authors" in a given "genre", and the portal to be the "reader" and interpreter of their "writings". (Goatly 2008: figure1, 1)

Returning to the question "where" to find the meaning of a query, Hirst (2006) describes three different views; the meaning can be in the *text* itself, in the *writer* of the text, or in the *reader* of the text.

If the meaning is in the text, cf the message as the focus of the poetic function of language (Jakobson 1960), there is no consideration of the writer, only the words in the text. In this case the meaning of a query could be defined as the description of the corresponding concept found in a terminology such as the Unified Medical Language System (UMLS) (NLM 2013a) for medicine. For instance, the semantics of the word "fever" is, according to the UMLS, "a documented body temperature higher than 38 degrees C., or 100.4 degrees F.", and we may say that the interaction is semantics driven. However, as for example the word "cold", a query may have several possible semantics, thereby to base a decision on the relevant answers to provide to the seeker may result in both incorrect and misleading conclusions at the receiving end. Similar problems would occur if the seeker would interpret the answer literally, without considering the answer abilities of the portal. For instance, if the seeker posted the query *sars* (Severe Acute Respiratory Syndrome) to 1177.se in the autumn of 2012, the answer set would be empty, even though there was a public interest in the disease.

If meaning is considered to be decided completely by the writer, hence a focus on the expressive function of language (Jakobson 1960), we need knowledge about the writer to interpret the text. With the information seeker as the writer the only thing we know is what is in the query. The seeker's goals and intentions are not known. Hence, normally it is impossible for the portal to adapt the answers to the writer.

The portal is the reader of a query, and if meaning is in the reader, emphasising the conative function of language (Jakobson 1960), then the query would mean whatever the portal thinks it means based on its background knowledge, and every document returned matching this interpretation would be the "right" one. Thereby, which documents were the "right" ones would be determined by the portal, not by the seeker.

As seen, there are several different ways to define meaning of queries and answers, but the only actual data at hand is the search log, i.e. saved infor-

mation on the interaction in terms of the query and the provided and chosen answers (possibly also with data on the seeker's location and used device), past experiences and, possibly, different semantic terminologies commonly used in life science and medicine. Hence, whatever way to try to interpret queries and answers by the seeker and portal, it is important that this also considers the risks of wrong or misleading communication, and that we can describe and analyse these choices in our conceptual framework to be able to fully understand the process of information search.

To summarise, for a portal to be able to answer a seeker's query and satisfy her needs in an expected way, the portal has to be able to "understand" the meaning or intentions of the seeker expressed by the query. However, an idealistic view where the meaning is found solely in the query itself will not consider aspects such as when the need was expressed, i.e. the seeker's query situation, tried to be addressed by a pragmatics view on language as an act of communication. Furthermore, by this communicative perspective the portal and seeker become reader and writer where the "need" is not solely obtainable from the query, but the actors on their own and the context of when and where the interaction takes place are all important when trying to decide the "right" answers to given queries. Or, in Jakobson's (1960) terminology, a conceptual model of health information search should account for both the addresser (information seeker), addressee (portal), message (query) and context, as well as the features of the contact (interface) and code of information search.

3.3 Models of information search

The challenge of describing information search, i.e. seeking and retrieving information, has been addressed by several researchers. According to Dinet, Chevalier and Tricot (2012), models of information search can be divided into *information science* oriented models, which describe information search behaviour, and *cognitive* models, describing mental processes involved in searching, where current models emphasise the *cognitive*, *affective* and *social* dimensions of human information search behaviour. However, a challenge with many of these models, as described by Ingwersen and Järvelin (2005), is the two paths' minimal integration. Hence, a concerted view on both the information seeking and retrieval, or a combination of cognitive and information science oriented models.

In a presentation of models of *information behaviour*, i.e. "the many ways in which humans interact with information" (Bates 2010: 2381), Wang (2011) provides a summary of several different models for *information seeking*, i.e. "the overall process in which an individual engages in order to satisfy a need to

bridge a knowledge gap in a problematic situation" (2011: 40), and *information retrieval*, i.e. the activities which takes place when a seeker retrieves information, possibly using an information system. The models are comprehensive and consider many different aspects such as the role of the *situation*, including features like cognitive capacity, affective state and information competency, and *contexts* as occupation, social role and demographics. As described by the author there are many different models, but in our opinion they all emphasise the role of the seeker at the expense of the importance of the information systems, e.g. a health information portal, and the cognitive aspects which are important, but not available to the system trying to satisfy the needs of the seekers. In our search games, we try to view the seeker and portal as equally important actors in an interaction, where the actual communication as expressed by the query and potential additional information is in focus.

In essence, these models of information search differ with respect to how they view the participants of the interaction, especially the role of mind (the context) versus matter (the query). To clarify, below is an example of a search log from 1177.se, of how Ann's interaction with the portal may have looked, and among the 112 answers presented by the portal she chooses the fourth, linked to a page on how to treat the symptoms in her query.[16]

```
2013-11-11:08-50-12, Göteborg, feber hosta, 4, 112,
Fakta-och-rad/Sjukdomar/Hosta-vid-forkylning/
```

Hence, the only thing we know in this case about the seeker and portal and their interactions is the query, context as encoded by the time and location, and answer as captured by the search log. The only information intentionally communicated by Ann is the query, and the context captures aspects she is aware of but probably not emphasising as descriptive of her needs, even though they may allow the portal to return better answers to her queries. In other words, the context and chosen answer may be seen as a glimpse of Ann's cognitive view of the problem. Hence the portal would benefit from considering both these implicit cognitive aspects as well as the explicitly communicated ones of the query.

In addition to the question of query versus context, many existing models focus on only the seeker as an actor, not the portal as an active part of the interactions, but today search engines are based on sophisticated analysis of seeker behaviours, hence they are as active participants as the seekers. Or, as described by a British newspaper on Google's latest development, called the Hummingbird, it "is paying more attention to each word in a query, ensuring that the whole query – the whole sentence or conversation or meaning – is

[16]The extract has been cleaned from some information captured by the log.

taken into account, rather than particular words" (The Guardian 2013). Still, the search engine has to assume the seeker to act *rationally*, that is, act to maximise the benefits with respect to the costs. In other words, the seeker tries to post a query which as well as possible reflects her information needs, based on her background knowledge and understanding of how the portal may react. Consequently, the portal tries to answer the query as well as possible, based on its understanding of the intentions of the query, and the search logs can be viewed as transcripts of rational activities.[17]

Notions like *rational actors* and the *sequential interaction* among them to achieve their goals lead us to the *theory of games*, but before turning our attention to this concept we will try to further motivate why it is relevant, based on *situation theory*, to model the interaction itself in addition to modelling the linguistic aspects.

3.4 An introduction to search games

Following the presentation above, semantics, pragmatics and the patterns of interactions, or discourse, are all important if we want to describe and analyse information search. At the core of the discussion we find the notion of *flow of information*, and with this as a starting point we end this chapter by briefly introducing our conceptual framework inspired by the theory of *equilibrium semantics* by Parikh (2010) and by Parfionov and Zapatrin's (2011) *game-theoretic* view of information search.

In the example of Ann, we can identify *entities* like Ann and the portal, their *properties* and *relations* among them, all of them viewed as *information* describing a *reality* of Ann interacting with a health portal to seek information. Each part of this reality, and its corresponding information, can then be said to correspond to a *situation*. For instance, one situation could correspond to the actual typing by Ann of the query "influenza" and another one to the answering by the portal. The situation of answering is related to the one of querying by the former said to be *constrained* by the latter and thereby (partly) defining the meaning, or *content*, of the situation. In the wording of information, we may say that the content is defined by the flow of information constraining the situations. This view underlies the work by Barwise and Perry in the 1980s on *situation theory* (Devlin 1991) of an informational universe containing individuals, relations, situations and constraints, aimed at providing a mathematical framework to describe real world situations.[18]

[17] It may be the case that the portal in addition to providing answers to the seeker's query would like to promote e.g. disease prevention or lifestyle changes.

[18] According to Devlin (2006) first presented in (Barwise and Perry 1983).

If we now try to tie all aspects discussed in this chapter and the example together, it can be depicted as in figure 6. When Ann passes the query "influenza" (φ_1) to the portal, she is either with probability ρ_1 in a situation s_1 where she is interested in treatment of the disease, or with probability ρ_2 in situation s_2 where she is interested in prevention of the disease.[19] In the first case after the query she will be in situation t_1 expecting treatment information, and in the latter case in t_2 expecting prevention information. However, the portal does not know in which one of these situations she is and consequently it does not know which type of answer she is interested in. In other words, it does not know whether to interpret Ann's query as a treatment query σ_1 or as a prevention query σ_2. Depending on the portal's interpretation, Ann may end up in several different situations. In situations v_2 and v_5 she has obtained the desired types of answers, but in v_3 and v_4 she has not. Moreover, this highlights the indeterminacy resulting from Ann not specifying her query. For instance, if she instead posted the query "influenza treatment" (φ_2) she would end up in situation v_1 with the expected answer from the portal. Similarly, if she posted the query "influenza prevention" (φ_3) she would end up in situation v_6.

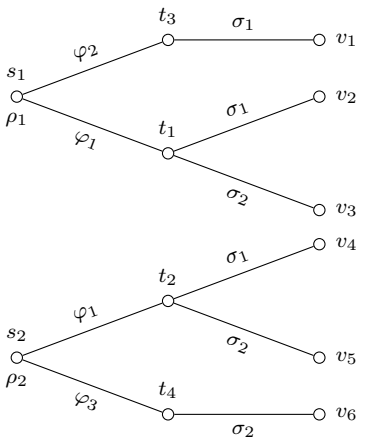


Figure 6: Game-theoretic view of communication.

Finally, when Ann's query and the portal's interpretation match in the sense that the answer satisfies Ann's needs, the system in figure 6 is in balance, or at *equilibrium*. Using Parikh's (2010) terminology, our model describes the *equilibrium semantics*, or *content*, of the interaction between Ann and the portal.

[19] Obviously we do not know Ann's situation or her potential interests, but we may add a situation "unknown" to the model with a probability which will make the "known" and "unknown" options sum to 100%. This assumption may be compared to the so called *closed world assumption* in logic reasoning (Reiter 1978).

Hence, the search log can be seen as a transcript of the interactions between information seekers and the health portal to obtain semantic equilibria.

The objective of a session between an information seeker and an information system can then be framed as a sequence of moves to achieve a semantic equilibrium. The notion of semantic equilibrium corresponds to an acceptable level of understanding between conversational parties to act or re-act in a feasible way, even in cases of potentially vague statements and need of clarification dialogues. For instance, Ann's query "influenza" resulted in a difficulty for the portal to decide the "right" move, thereby in a risk of not achieving a semantic equilibrium. However, if she had instead posted the query "influenza treatment", it would have resulted in an equilibrium with the portal understanding Ann's needs, acting in the "right" way in her eyes and she would consider the portal to be trustworthy and competent. Hence, semantic equilibria in the context of information search is related to the concept of trust and its consequences.

As discussed in the previous section, there are many different ways to describe, or model, information search. Our search games share several features of these, like trying to capture aspects related to the social dimension of searching where the seeker may act on her own as well as part of a group. But also ones related to emotional and affective factors influencing, or being influenced by, seekers' trust and dependency on a health portal to address their needs and knowledge in the field of medicine. However, a feature of our model which stands out by its foundation in game theory is its, theoretical, ability to provide a basis for executable models for predictive analysis and to automatically induced models, or preference relations, based on search log data. Hence, the model does not only provide a way to "depict" existing search behaviours, but also to facilitate what-if analysis of changed portal and and seeker behaviours.

By this introduction we have tried to describe why linguistic aspects like semantics and pragmatics, but also interaction patterns per se, are important when we want to establish a conceptual framework for information search. We have also introduced the cornerstones of game and situation theories, which will be used as a basis for this work. In the following chapters we recapitulate the most relevant aspects of these two theories before we finally end up with our conceptual framework.

4 GAME THEORY PRIMER

In *Zur Theorie der Gesellschaftsspiele* (von Neumann 1928: 295) the author wrote: "any event – given the external conditions and the participants in the situation (provided that latter are acting of their own free will) – may be regarded as a game of strategy if one looks at the effect it has on the participants" (English translation in (Tucker, Luce and Kuhn 1959: 13)).

John von Neumann (1903–1957) was a Hungarian mathematician considered to be one of the founding fathers of *game theory* once aimed to revolutionise the early twentieth century view on economics and social theory. According to Leonard (1995) several different paths of challenges led to the publication by von Neumann and Morgenstern of the "Theory of Games" (1944).[20] Firstly, the turn of the century 1900 and the discovery of paradoxes in mathematical set theory had led to a crisis in the foundations of mathematics, especially when there was a trend to try to describe science *systematically* and where the description itself could lead to insights by *abstract reasoning*. This *structuralism* targeted areas reaching from physics and linguistics to economics. In other words, mathematics was seen as the *language of science* and, according to its proponents, as many different phenomena as possible should be explained by mathematics. Secondly, von Neumann shared the interest with many other mathematicians, e.g. Zermelo (1912), to study the mathematics of parlour games such as chess, "matching pennies" and "paper, stone, scissors". Hence, mathematics became a way to *model human behaviour*, especially among "rational" players.

At the same time the economist Oskar Morgenstern (1902–1977) criticised the current formalisation of economics and its use of statistical and mathematical methods, which according to Morgenstern ignored the knowledge and behaviour of the involved parties of economical transactions. Morgenstern was not a mathematician, but more and more he saw the need for a new type of mathematics in economic theory taking into account the *processes of interac-*

[20] A more up to date presentation of the history of game theory is presented in (Leonard 2010).

tions. Morgenstern and von Neumann met in the late 1930s when they had both left Europe for America, sharing the interest to outline a *theory* for *rational behaviour*. This collaboration led to the *theory of games* today applied in a wide variety of fields from auctions and contracts to political science, evolutionary biology and linguistics.

Trying to define game theory is trivial, at the same time as almost impossible. Myerson (1991: 1) defines it as "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers", where these models, called games, are well-defined in terms of set theory. Hence, with formal set-theoretic definitions of the players and the rules of the game, allow for discussions on notions related to how the game is, or should be, played to achieve the aims of the players. Still, to understand the birth of game theory, its role and differences in comparison to other points of view, one has to study the existing view of economics and behaviour during the first half of the twentieth century (see (Leonard 2010) for an in-depth presentation). However, Aumann (2008) calls game theory an "interactive decision theory", which in our opinion describes its definition and role – it aims at providing a well-defined framework to study *decisions*, their *prerequisites* and *consequences*, which are interactive by their nature. If we then apply this perspective to information search, we may say that we would like to establish a framework to study seekers' and portals' decisions regarding queries and answers to achieve a feasible interaction and outcome for both parties.

At the core of game theory we find the question of how to make decisions based on knowledge and assumptions of the surrounding world, and in the rest of this chapter we begin in section 4.1 by presenting a brief overview of the notions of *preference* and *utility* and how these lead to concepts like *rational choice* and *trust* (section 4.2). Then we present the theory of games (section 4.3) focusing on notions of relevance for our research (section 4.4). Finally, we discuss how game theory allows for abstract analysis of specific games, but also some of the limitations with a game-theoretic view of the world (section 4.5) and how some of these can be resolved by incorporating the notion of *situation* in the models (section 4.6).

4.1 Preferences and rationality

To introduce the basics of game theory we begin by presenting an example of a game called "paper, stone, scissors", abbreviated PSS. In this simple game both players[21] simultaneously choose between three alternatives, or *moves*, called "paper", "stone" and "scissors", where paper beats stone, stone beats scissors,

[21] We will use the terms *player* and *actor* interchangeably in this work.

and scissors beat paper as seen in figure 7. For instance, if A believes that B is going to choose paper, A's preferred choice would be scissors and not stone or paper, since scissors beats paper but paper beats stone.

		B		
		Paper	Stone	Scissors
A	Paper	Draw, Draw	Win, Loss	Loss, Win
	Stone	Loss, Win	Draw, Draw	Win, Loss
	Scissors	Win, Loss	Loss, Win	Draw, Draw

Figure 7: Preferences in the PSS game for players A and B.

At the core of the above game we find the notion of *preference*. For instance if A thinks B is going to choose stone A's best *strategy*, or sequence of moves, is to choose paper, which encodes the "relative merits of any two outcomes for the player with respect to some criterion" (Slantchev 2007: 2). Hence, the preference relation can be defined as a (finite) binary relation $\succ$ on the *outcomes* $\mathcal{O}$ of the possible (combined) strategies of the players of the game.[22] By $x \succ y$ we mean that (the outcome of) strategy x is strictly preferred to (the outcome of) y. If x is preferred to y or the player is indifferent between the outcomes we denote it $x \succeq y$, and $x \sim y$ in the case of indifference. For instance, in PSS player A chooses scissors and B prefers paper is an outcome, in this case with A as the "winner". If B's choice was based on that she assumed A was going to play stone, we had (stone, paper) $\succ$ (stone, scissors) for player B. The notion of preference is fundamental in game theory, but let us take an informal view on this notion before proceeding with formal definitions.

Game theory was described as "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers" (Myerson 1991: 1), with the two keywords *conflict* and *cooperation*. However, a prerequisite for conflict is the ability to make decisions, or *choices*. These in turn require that we are able to compare alternatives, hence some type of preference relation among outcomes is needed. Instead of comparing the outcomes, we may compare the paths, or strategies, achieving the outcomes, assuming that following a path would lead to given outcome. Hence, the notion of preferences among moves ends up as crucial to achieve the aims of game theory.

Another important aspect is that we assume the actors responsible for making these choices to be "intelligent" and "rational", and what we in common

[22] Formally, the set of outcomes is defined as a function o from a cross-product of sets of moves to a set of values which allows comparisons of strategies. Since it is not needed in the rest of the presentation, the formal mathematical definition is omitted. Moreover, instead of referring to comparing outcomes $o(x)$ and $o(y)$ of strategies x and y, we will simplify the presentation by alluding to preferences among strategies.

language mean by these two terms may differ. However, if we make use of the parlour game metaphor, we do not get involved in a game if we do not want to "win", or at least play as well as we can. Hence, we want to make the "right" choices along the unfolding of the game, and since we said the notion of preference captures the idea behind "choice", we try to define properties of this relation which would be in line with "rational" behaviour in a game.

As in the work by Slantchev (2007), we introduce the following assumptions for the preference relation, helping us to define the prerequisite in game theory of rational behaviour of the players.

Assumption 1. *Preferences are **asymmetric**, i.e. for no pair x and y of strategies both $x \succ y$ and $y \succ x$.*

The first assumption defines that among any two strategies an actor has to be able to decide which one she prefers, or that she is indifferent or ignorant about the choices. For instance, if a PSS player has to choose between stone and scissors, given that the other player is assumed to play paper, not both these alternatives are better than the other one. If we were to capture the idea of a "rational" player, hopefully we can agree that a player who is not able to make up her mind on different choices is most probably better off not entering the game.

Assumption 2. *Preferences are **negatively transitive**, i.e. if $x \succ y$, then for any strategy z either $x \succ z$ or $z \succ y$ or both.*

The second assumption means that for any new potential move, the actor has to be able to relate its strategy to already established preferences. In the PSS example we have, for instance, if player A prefers paper to scissors whenever B is expected to choose stone, then either she has to prefer paper to stone or stone to scissors or both. As in the case of a so called asymmetric preference relation, a "rational" player who is not able to relate a potential move to her existing preferences, will run into problems in the game.

As proven by Slantchev (2007), these two assumptions imply that for no strategy x we have $x \succ x$ (*irreflexivity*), if $x \succ y$ and $y \succ z$ then $x \succ z$ (*transitivity*), and if for a finite given integer n, $x_1 \succ x_2, x_2 \succ x_3, ..., x_{n-1} \succ x_n$, then $x_n \neq x_1$ (*acyclicity*). It is important to remember that these properties have to hold for strategies, hence moves given the history of previous moves of all players. In other words, stone beats scissors in PSS does not imply that every strategy with a choice of stone should be preferred to one with choosing scissors. In common language this means that you cannot prefer a move to itself (reflexivity), you need to be able to "chain" paths of preferences in a sound way and that circular justifications are not "rational".

These two assumptions of asymmetric and negative transitive preferences lead to a number of natural consequences reflecting our intuition of a relation which aims to work as the basis of a formal definition of rationality.

Proposition 1. *If the preference relation $\succ$ is asymmetric and negatively transitive, then the following hold:*

1. *The preference relation $\succeq$ is complete, i.e. for all distinct strategies x and y, either $x \succeq y$ or $y \succeq x$ or both*
2. *The preference relation $\succeq$ is transitive, i.e. if $x \succeq y$ and $y \succeq z$, then $x \succeq z$*
3. *The indifference relation $\sim$ is reflexive, i.e. for all strategies x, $x \sim x$*
4. *The indifference relation $\sim$ is symmetric, i.e. for all strategies x and y, $x \sim y$ implies $y \sim x$*
5. *The indifference relation $\sim$ is transitive, i.e. if $x \sim y$ and $y \sim z$, then $x \sim z$*
6. *If $w \sim x$, $x \succ y$ and $y \sim z$, then $w \succ y$ and $x \succ z$*

Proof. See Slantchev (2007). □

The first two properties say that if a player can carry out a line of reasoning based on the preference relation that satisfies our two basic assumptions, then she will be able to make decisions on all possible pairs of choices and no paths of preferences will lead to inconsistencies or circular reasoning.

We are now ready to formally define the notion of *rationality*, meaning that given any two alternative strategies an actor can determine whether she likes one at least as much as the other (*completeness*) and no sequence of pairwise choices will result in a cycle (*transitivity*). We say that the preference relation $\succeq$ is *rational* if it is complete and transitive (Slantchev 2007). In plain English, the actor is able to make up her mind about any two alternatives if one is better than the other, or if they are equally good, and she will not end up in circular preferences with strategy x being better than y and y being better than z, but z being better than x.

Even though the preference relation suits the needs of a game, in many cases it is easier to use numerical values, called *utilities*, to rank alternative strategies with respect to their preference. A function $u : \mathcal{O} \to \mathbb{R}$ is a utility, or *payoff*, function representing preference relation $\succeq$ if for all strategies x and y we have that $x \succeq y$ if and only if $u(x) \geq u(y)$. It is important to note that utility only specifies the *ranking* of alternatives and not how "much more" one

strategy is preferred to another. Finally, let the *best* strategies among a set of alternatives be the ones that have the maximum utility.

The concept of utility discussed by von Neumann and Morgenstern (1944: chapter 3) is considered fundamental in game theory and economics as described by Marshall (1920: 78):

> "Utility is taken to be correlative to Desire or Want. It has been already argued that desires cannot be measured directly, but only indirectly, by the outward phenomena to which they give rise: and that in those cases with which economics is chiefly concerned the measure is found in the price which a person is willing to pay for the fulfilment or satisfaction of his desire."

Returning to our PSS example, we have that the payoff of A to play stone, when B is assumed to play paper, is less than to play scissors, figure 8. Or mathematically, for player A we have $(\text{scissors}, \text{paper}) \succ (\text{stone}, \text{paper})$ since $1 > -1$, or A is more willing to pay to play scissors than stone, assuming she knows what B is going to play, to feel the pleasure of winning the game.

		B		
		Paper	Stone	Scissors
A	Paper	0, 0	1, -1	-1, 1
	Stone	-1, 1	0, 0	1, -1
	Scissors	1, -1	-1, 1	0, 0

Figure 8: Payoffs in the PSS game for players A and B.

The following proposition says that if we define the preference relation in a game as above, then and only then it is possible to encode it by comparing pairs of numbers with the ordinary greater-than-or-equal-to relation, i.e. the payoffs of choices.

Proposition 2. *A preference relation can be represented by a utility function only if it is rational.*

Proof. See Slantchev (2007). □

In other words, if we accept the assumptions of a preference relation and would like to express it in terms of the established notion of utility, then we must require it to be rational. An even stronger result is obtained if we just assume the set of outcomes to be finite.

Proposition 3. *If the outcomes on which the preference relation is defined is finite, then the relation admits a numerical representation if and only if it is asymmetric and negatively transitive.*

Proof. See Slantchev (2007). □

In other words, we can use the numerical inequality and equality relations to capture our intentions of rationality, that is, any rational player should be able to "numerically" justify her strategies.

In our example, each player had to make their moves based on no knowledge of previous moves by either party in the game. However, in many real life situations the preferred actions change over time based on the progress up to the situation at hand. For instance, in a search scenario the information seeker has some knowledge I_0 of the portal's "behaviour" when she posts her query to the search engine. She then receives some answers which may change her knowledge to I_1, when posting the next query. Hence, we extend the notion of preference to strategy x is preferred to y by player i given information I_j, denoted $x \succeq_i^j y$. However, this provides some challenges with respect to the notion of rationality and we define the notion of *bounded rationality* for player i with respect to the information I_j as the relation $\succeq_i^j$ is complete and transitive.[23] Consequently, we call player i's behaviour in a game *rational*, if for all played strategies $j = 1, ..., m$ we have that the relation $\succeq_i^j$ was bounded rational, and that the game is rational if all players' behaviours are rational.

The concept of *bounded rationality*, introduced by Simon (1997) is commonly used in economics to denote the impossibility of a "rational" person being globally rational: "Global rationality, the rationality of neoclassical theory, assumes that the decision maker has a comprehensive, consistent utility function, knows all the alternatives that are available for choice, can compute the expected value of utility associated with each alternative, and chooses the alternative that maximizes expected utility. Bounded rationality, a rationality that is consistent with our knowledge of actual human choice behavior, assumes that the decision maker must search for alternatives, has egregiously incomplete and inaccurate knowledge about the consequences of actions, and chooses actions that are expected to be satisfactory (attain targets while satisfying constraints)." (Simon 1997: 17). As discussed by Barros (2010), both the notion of bounded rationality and the related *procedural rationality* introduced by Simon are rather vague. However, from Barros' perspective the procedural rationality, based on the notions of search and *satisficing*, is *interpreted* in terms of algorithms and computability. Hence, bounded rationality reflect what is possible to "compute" with given knowledge and (mental) capacity. However, this extended notion of preference, and its associated rationality, leads to fundamental problems to game theory in its classical sense, but as discussed in

[23] Since we consider the information seeker's and portal's behaviour to be consistent with the line of reasoning introduced by Simon (1997), and capture our intentions, we re-use the phrase in this work.

section 4.5 in a way of less importance to our use of the theory.

To summarise, following the work by Slantchev (2007) on rationality and by the addition of bounded rationality, we call a player rational if she is able to decide on her alternatives, based on her knowledge at hand, in each step of the game, and the justifications in each step are not circular.

4.2 Trust

In the introduction of this thesis a fundamental concept, possibly not explicitly stated, in the interaction between health information seekers and a portal is the notion of *trust*. For instance, the seeker trusts the portal, or considers it to be *trustworthy*, to provide adequate information and advice on preventions, treatments and care providers. However, the notion of trust, and its prerequisites and consequences, is a non-trivial topic.

As stated by McKnight and Chervany (2001: 28), "[t]rust and mistrust are widely acknowledged to be important or even vital in cooperative efforts in all aspects of life", and a search play is definitely one type of cooperation where even the health of an information seeker may rely on the trust in the portal. However, trust can be defined as a noun, verb, a personality trait, belief, a social structure and behavioural intention. Moreover, it may be characterised in terms of the bases by which it forms, e.g. as outcome of a process, trust in an institution or in terms of recognition of affirmative aspects or knowledge, and we may even trust and distrust at the same time and social mechanism to deal with risk or uncertainty.[24] Hence, following McKnight and Chervany (2001) efforts to describe a typology of trust we will try to characterise it in our setting with two rational actors with intentions of collaboration to achieve the expectations of the seeker.

We begin by noting that trust can originate in a general *disposition to trust* reflecting "the extent which one displays a consistent tendency to be willing to depend on general others across a broad spectrum of situations and persons" (McKnight and Chervany 2001: 38), and divided into the subconstructs of faith in humanity and a *trusting stance*. Since we view health portals to be actors in a similar way as an information seeker, we can paraphrase these constructs in *faith in portals* and view that portals are "well-meaning and reliable" (2001: 39).

The disposition to trust impacts to which degree one "believes, with feelings of relative security, that favorable conditions are in place that are con-

[24] In our context we let *distrust* denote the "opposite" of trust, that is, the less you trust the more you distrust. However, distrust can be treated in a similar topological way as trust, see (McKnight and Chervany 2001) for further details.

ducive to situational success in a risky endeavor or aspect of one's life" (2001: 37), hence an *institutional* trust. In other words, if an information seeker believes there is a *structural assurance* like regulations, processes, procedures etc in place to ensure the portal provides adequate health support and that the seekers' concerns and needs are "normal" for seekers, i.e. a high *situational normality*. The disposition to trust also affects "the extent to which one believes, with feelings of relative security, that the other person has characteristics beneficial to one" (2001: 36), hence *trusting beliefs*. In other words, the trust is not situation-, but actor-specific and reflects to which degree the seeker believes in the portal's *competence*, that its aim is to make "good" and not "profit" (benevolence), that its answers are "true" (integrity) and that its behaviour is *predictable*. Finally, the disposition to trust impacts the *trusting intentions*, i.e. if the seeker is "willing to depend, or intends to depend, on the other party with a feeling of relative security, in spite of lack of control over that party, and even though negative consequences are possible" (2001: 34). Hence, it consists of the *willingness to depend* and the "extent to which one forecasts or predicts that one will depend on the other person, with a feeling of relative security" (2001: 34), called the *subjective probability of depending*. The trusting beliefs and intentions both describe "interpersonal" relations between the seeker and portal, and both impacts the seeker's behaviour when using the portal, hence her *trust-related behaviour* considering *cooperation* viewing the interaction as a process, *information sharing* by being willing to share potentially vulnerable information, *informal agreement* to both seeker and portal to do their "best", *reduced control* and trust of the provided portal answers, acceptance of *influence* and *decision-making power* of the portal to "follow" given advice and the interaction, or *transaction*, to be carried out in good faith and ethics.

To summarise, seeker's trust in a portal can be defined, and being impacted by, her disposition, view of the portal as an "institution", beliefs, intentions and behaviour, and the main way for the portal to achieve and maintain the seeker's trust is by its treatment of her queries and the answers it provides.

If we now turn our interest to trust in the context of our search games, we have that the notion of rationality is closely related to notions like predictability and trust, by a rational behaviour leading to a higher degree of predictability by the other players of the rational actor's moves and preferences. This in turn, in collaborative games, results in a basis for trust, and less need for "irrational" actions by the other players.[25]

[25] In the case of parlour games, this predictable rationality in the eyes of the competitors may be an unwanted feature, and a player may act rationally, but try to give the impression of irrationality or a different rationality than intended.

4.3 The game

In section 4.1 we defined the notions of preference and rationality, which we claimed were the core of any concept of a "game". However, we have not yet formally defined this concept, but before doing so it is important to emphasise that one of the philosophical drivers of the creators of game theory was to introduce mathematical rigour, with a basis in set theory, into the discussions on parlour games as well as economics in the 1930s. As a consequence, the theory may seem overwhelming, but keep in mind that it is only aimed to capture our intention of a "game" in a well-defined way and, as will be seen, in our context the notions will be simplified to increase readability and focus on the important aspects.

Let $N = \{1, ..., n\}$ denote the set of *players,* or *actors*. Then we assume that for each player i there is a set $\mathcal{A}_i$ of all *actions*, or *moves*, that can be chosen by her. A *round* is an n-tuple $(a_1, ..., a_n)$ of actions where each action a_i is chosen among the player's available actions $\mathcal{A}_i$.

In our PSS game we have two players $N = \{1, 2\}$ with the same sets of moves, $\mathcal{A}_1 = \mathcal{A}_2 = \{\text{paper}, \text{stone}, \text{scissors}\}$, where $(\text{stone}, \text{paper})$ and $(\text{scissors}, \text{paper})$ are examples of rounds. The latter round says that player one played scissors and player two preferred to play paper.

A *strategy* x_i^j of a player i at time point $j \geq 1$ is defined as a function from the player's information I_i^{j-1}, i.e. her background knowledge including the played rounds at time points $k < j$, to an action in $\mathcal{A}_i$.

A (strategic) *game* is a set of rational players, their action sets and preferences for strategies. A *play* of length m of the game is defined as an n-tuple $(x_1, ..., x_n)$, where $x_i = (x_i^1, ..., x_i^m)$. In other words, a play of a game describes a sequence of rounds consisting of the players' moves, all acting rationally at their respective moves in a round. This can also be expressed as each player chooses her "best" strategy knowing her own past moves, preferences and the other players' past moves.[26]

To exemplify our definitions, let us look at a variant of the PSS game, called PSS-2, where each game consists of two sequential rounds based on the same preferences and payoffs as in the case of the PSS game. Moreover, in this game the first player makes her move by writing it on a piece of paper, then the second player does the same and finally they show each other the papers to decide the outcome of the round. The one who wins the last round is considered the

[26]We would like to emphasise that our definition of a game may be seen as a layman way of saying that in a game you do your "best" given your knowledge. For discussions on mathematical definitions of "best" and other aspects, we refer to, for instance, seminal work by von Neumann and Morgenstern (1944) or the concise introduction by Leyton-Brown and Shoham (2008).

winner of the play. In this game, each player tries to decide her move based on an interpretation of earlier plays of the game. For instance, if the first player, called Paper Boy, has heard that his antagonist, Rock Star, is known for choosing stone in the first round, then it may be a good idea to start with a paper move. Hence, Paper Boy starts in a situation s having heard this rumour or in one being ignorant of this rumour. Depending on if Paper Boy decides to make use of the information, he ends up in different situations (t, t') following a paper move or not, respectively. Then it is Rock Star's turn to play along these preferences, but in this case he has heard a rumour of the first player to be a player who starts with a paper move. Hence, he plays scissors and ends up in situation u. Following Rock Star's move it is time for Paper Boy again, remembering the last round of the game, for instance, choosing paper again to end up in situation v, and Rock Star to end with a scissors move leading to the final situation w. In this game the setup is the same as in the PSS game, with preferences and payoffs per round as in figures 7 and 8. The difference is that in this case it is played in two rounds and a play can evolve as in figure 9.

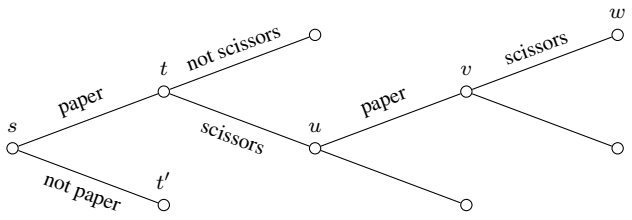


Figure 9: Game-theoretic view of the PSS-2 game.

In this play, Paper Boy's (player 1) information at the beginning of the game ($j = 1$) was that Rock Star (player 2) tends to play stone. Therefore he chose the move paper based on the information I_1^0 since it complies with his preferences and the assumption that Rock Star will play stone in the next move. Similarly, Rock Star's information I_2^0 that Paper Boy often plays paper makes him choose the strategy of playing scissors which complies with his preferences. Since both players stick to what they know about their competitor, the next moves will take place in a similar way. By the rules of the game, this turns out to be fruitful for Rock Star since his strategy (scissors, scissors) beats Paper Boy's (paper, paper) due to (scissors) $\succ_2^2$ (paper), i.e. the payoff for scissors is better than paper whenever the competitor plays paper.

From parlour games we know that not all plays of a game are equally good according to the players, hence some are called "wins", some "losses" or "ties". These properties of games, or plays, with a focus on our games of interest, is the topic of the next section. However, before proceeding it is impor-

tant to stress that the definition of a game does not include any requirements of a notion of win or loss, but only that its actors act in a rational way. The last aspect is often associated with parlour games, but not even these require rational behaviour. Hence, the term "game" might be misleading, but as discussed in the introduction of the chapter, originates from the early interest in the mathematics of chess.

4.4 Properties of games

In the previous section we introduced the notion of games and the concepts of actions and strategies, and comparison of these by their utilities, and the next step is to describe some important types of games and properties of these.

One of the simplest types of games is *static games*, e.g. the PSS example, where every player only performs one action simultaneously. In a static game each play consists of only one round, otherwise it is called a *dynamic game*. If the players make their moves in a given order, the game is called *sequential*. In a static game, or one with known round, we may omit the round indexes from our formalism. For instance, in the PSS example, if I_1 contains the assumption that player two is known to play scissors, then $(\text{stone}) >_1 (\text{paper})$. However, as will be presented in chapter 5, our main interest is in sequential dynamic games, where the actors consist of an information seeker, or a group of seekers, and the underlying search engine of a (health) portal. We call these games *search games*, and they are formally defined in the next chapter.

Up until now we have not discussed the notions of *wins* and *losses*. In classical game theory these are related to the ability to compare the players' utilities, and we may say that a player has won a play if her total pay-off is greater than the opponents' when the play is ended. The term "ended" may itself be defined in terms of payoffs, or just by defining a number of rounds to make up a play. One way to combine these definitions is to say that a player has *won* the play if the opponent is not able to make any sequence of moves to obtain the greatest pay-off. Applying these concepts in our search games lead to four different possible outcomes: *win-win*, *win-lose*, *lose-win* and *lose-lose* for the seeker and portal, and we will discuss each of them in some detail.[27]

First we have to define the notions of win and lose for the seeker and portal respectively. Basically, the more "happy" the seeker is with the provided answers, the more of a "winner" she probably feels. Seeker happiness may also be defined as the ease with which she can find the relevant answer, or click,

[27] Since the aim is to introduce game theory in the context of information search, we will exemplify the preceding concepts in this type of games. However, the notions also hold for games in general.

among the provided ones. Thereby, the easier an adequate click can be found, the more of a winner, or the higher payoff she has obtained. However, it is important to remember that it is only the preference among the alternatives which counts and not any numerical differences. In other words, a seeker may be more happy with answer A than B, but it is not possible to provide a numeric value as a "measure" of this difference.

A similar reasoning can be used to define the notion of win for the portal, and we have that both the seeker and portal aim for wins. The best possible play would then be one with only one query and one answer highly appreciated by the seeker. However, in reality the efforts of the actors in a search game aim at obtaining and providing "good enough" answers, and they may even "cooperate" to avoid some of the unwanted outcomes. Figure 10 exemplifies the different outcomes of a search play. Lose-lose plays are, for instance, ones where the seeker provides an "adequate" query, but the portal cannot provide any answers, hence the play is "incomplete". Obviously, both parties would like to avoid this type of play. The lose-win games are interesting, since in this case the portal considers itself to have provided "good enough" answers, but still the seeker is unhappy. This may result when the seeker's query was wrongly interpreted by the portal.

		Portal	
		Win	Lose
Seeker	Win	Seeker obtains answer on flu treatment when posting query with flu symptoms	Seeker clicks answer on hangover cures, but portal had preferred click on risks with alcohol intake
	Lose	Seeker obtains no answers due to misspelt query	Seeker obtains no answers to the query 'smallpox' since the portal considers the disease to be extinct

Figure 10: Examples of search outcomes for *seeker* and *portal*.

Another aspect of games, also relevant in the context of search games, considers the *degree of cooperation* among the players and reflects if and how "coalitions" and "agreements" are possible among the players. In *strictly competitive* games no cooperation is possible, and in terms of payoffs if one player "gains" a certain profit the other players will have to "lose" a similar amount. Hence, the payoffs of the players have to sum up to zero for each strategy. These games are therefore called *zero sum games*. At the other extreme of cooperation, we find games of *pure coordination*. In these games all players will win or lose the same amount, thereby the players can be called *partners.* Based

on these notions, we assume each search game to be cooperative, since it is in the interest of both parties to help each other achieve their goal.

Every game has the aim of reaching an end point, for instance one where no player has the interest or ability to play another round of the game or change their chosen strategy, and this point in the game is called an *equilibrium*. An equilibrium does not imply that the play was a win-win, only that it has come to an end.[28]

4.5 Limitations of games

Traditional game theory in its simplicity obviously has limits, for instance assuming the players to be rational and goal directed. Hence, it excludes changes in players' behaviour due to internal and external aspects like threats, promises and persuasions, but also other players' rational arguments if we do not allow the players to consider the previous rounds of a game and their background knowledge of the given types of games. Based on the work by Howard (1994a, b) we may say that "life is a drama, not a game", and a similar concern is raised by Allott (2006) in the context of communication since it also assumes that the players share a common knowledge of the game.

In chapter 3 we also noticed that several different aspects such as background knowledge and expectations may be involved in the interpretation of actions like communication or information search. Thereby one may argue that a game-theoretic model of information search is too simplistic, but then one has to keep in mind that the only information we have is what is captured by the search logs, and as will be seen in the next section, this makes up the information, or knowledge, alluded to in our discussions above. Hence, we extend our games with this limited view of situation-based decision making, thereby accounting for the needs in the context of information search.

4.6 Situation-anchored game theory

Before we continue with our presentation of search games, it is important to reflect on the specific type of games we are interested in – sequential dynamic games with two players and bounded rationality. This type of games lay beyond the ones introduced by von Neumann and Morgenstern almost 80 years

[28]The concept of equilibrium is well-known in game theory, especially so called *Nash equilibria* after John Forbes Nash Jr (Nobel laureate 1994 and the inspiration for the film *A beautiful mind* in 2001). Nash proved in 1951 that if each player has chosen a strategy and no player can benefit from changing strategies while the other players keep theirs unchanged, then the current set of strategy choices and the corresponding payoffs constitute a Nash equilibrium.

ago, especially allowing the rules, i.e. the preferences, of the game to change over time due to the bounded rationality of the players. Another way to express this is to say that the game should be *anchored* in the current situation.

As described by Parikh (2010: 9), the ability to represent the world to ourselves and communicate it to others, especially through language, is what makes us human. This allows us to describe the *reality*, or the world, as a space, called *information*, of entities. This space contains *individuals* having *properties* and standing in *relations* to each other. Collections of such entities make up what we call *situations*, and are the basis for our view of ourselves and others, as well as a foundation for our decisions regarding the future. Since these collections of individuals, their properties and relations are fundamental they have been given the name *infons* to represent "states of affairs", or *facts* when true of some existing situation (Parikh 2010: 38–39).

The space of infons, and their mathematical theory, was first developed by Barwise (1989), see also the work by Devlin (1991) on logic and information, under the umbrella of *situation theory*. Two key features of this theory are that an individual seldom has full information about a situation and has to act on partial information, and an intention to differentiate between language and meaning. The latter means that the theory tries to differentiate between, for instance, a property and the objects in a situation with that property or the type of an object and the object itself.

Before drowning in the theory of situations, its philosophy and mathematics, we note that our introduced game theory lacks the notion of situation even though it is based on the notion of preference which assumes some kind of ability for players to act based on their information on their "situations". Considering the history and application of game theory, this may be understandable, since the actors, as in parlour games, knew the rules of the game and (in theory) the strategies towards success and the consideration of situations boiled down to choosing the right strategy based on the other actors' preferences and perceived opportunities. Hence, the actors are not capable of "changing their minds" as a result of "negotiations" in the sense of, for instance, threats, promises, persuasion and rational arguments (Howard 1994a: 187–189).

In the light of our interest in a game-theoretic model of information search, we may thereby claim the model to be inadequate since information seekers are not robots acting in a predetermined way, even though the portal may act in a determined fashion.

To summarise, on the one hand game theory allows an intuitive and formal way to describe interactions of players based on notions like preference and rationality. However, it does not allow realisation or consideration of concepts like change and development. There are different ways to approach this challenge, for instance, as described by Howard (1994a) in his *drama theory*,

which can be seen as an extension of game theory into a "soft" version where games are allowed to be "transformable" over time. Another approach taken by Parikh (2010) in his *equilibrium semantics* as a mathematical framework of meaning for natural language is to combine the theories of games and situations in a theory of semantics. Both these approaches are feasible also in our setting, but with a risk of losing the intentions, and ourselves, in a mathematical or philosophical machinery. Hence, we will in the next section discuss the information at hand in information search and the consequences from a game-theoretic perspective.

4.6.1 Information, situations and search

When an information seeker posts a query to an information portal, the only data available to the portal is the query, potential context information and possibly records of past searches. The context information may let the portal know where the seeker is located, if the interaction takes place with a mobile device and the time of the interaction. Hence, the context is a realisation of the infon describing the seeker's situation at a level of abstraction available to the portal. The query and the context is the only information available to the portal when trying to figure out the needs of the seeker. Moreover, the overwhelming majority of today's interactions consist of only one query and set of answers among which the seeker chooses one or more to gain knowledge. Hence, there is no development or change of the interaction taking place in our setting as alluded to by Howard (1994a) as a problem with game theory in general.

Another topic related to the notion of information is if the actors have *perfect information*, i.e. if they correctly "remember" their past moves in the interactions, and if it is *complete*, i.e. both actors know all payoffs and strategies available to the other actor. Considering information search, the portal registers the seeker's queries and context and, assuming it keeps track of the "address" of the seeker, we may consider it to, at least in theory, be able to recapture both the seeker's and its own moves in the play. However, even though we expect the seeker to remember her queries and to know the associated context, it is questionable if she is able to recall the portal's move made up of an ordered list of answers. Still, we may assume she remembers the position of the, possibly, chosen answer in the list and the information it contained. Moreover, we expect she remembers the gist of the answers preceding the one she chose, and we could define the portal's move to consist of this list of answers ending with the chosen one. If so, we may claim the seeker to have perfect information of the portal's moves of the play. The notion of completeness assumes the actors to be fully aware of the other actors' available strategies, and for search games

this is not the case. Hence, we consider our intended games to be ones where the actors have perfect, but not complete, information of the other actor.

4.6.2 Situation as query

If we study the type of information captured by the context of a query, it is in most cases related to location, time or device, and used by portals to "filter" answers to satisfy constraints with respect to location, time or the way of presenting the answers on mobile devices. Hence, in the first two cases the context can be viewed as just two additional query terms to be accounted for by the information portal. Consequently, for certain types of realities as in our case, the situations and their use in models may be described as just an additional component of an actor's move. Thereby, no extension along the lines of drama theory (Howard 1994a, b) or equilibrium semantics (Parikh 2010) are necessary to establish a game-theoretic framework for information search, still we will call the underlying game theory situation-anchored to emphasise that we account for this type of information in our models.

Treating context information as just an additional set of query words can be argued to be incorrect, since the query words typed, or chosen, by an information seeker are intentionally passed to the portal, but parameters like used device, time and location may be set by the portal without the seeker being aware of it. This topic is discussed in section 9.2.3, where we conclude that existing portal solutions, where seekers are unaware of the additional information used by the portal as part of the "interpretation" of a seeker's needs, are not desirable and context information to be treated as intentionally provided by an information seeker.

5 SEARCH AS A GAME

In the last chapter we introduced *situation-anchored game theory* as a framework to describe interactions among actors whenever the situation, or context, impacts the preference among moves. In this chapter we present how this framework can be used to model the interaction between information seekers and information portals.[29]

In addition to Parfionov and Zapatrin's (2011) game-theoretic view of information search, we are inspired by Parikh's (2010) model of semantics, describing communication as a situation-anchored sense-making game with an actor trying to convey a message by utterances hopefully interpreted, and acted upon, in the intended way by the recipient. These two paths of research are briefly presented in section 5.3.

Information search as carried out by trying to find meaningful information on a health portal to address given queries may take place by several different means. For instance, browsing concepts to identify relevant "anchors", or typing queries and choosing among the suggested answers, linking to detailed information. At the core of both these approaches we have the notions of *sessions* of *interactions* among an information *seeker* and a *portal* expressed as (implicit) *queries* and *answers* guiding the information seeker (section 5.1). As discussed in chapter 3 there are several different ways to describe interactions with queries and answers in information search, and *game theory*, as outlined in the previous chapter, is one approach. In our opinion, it is well suited for our needs to view search as an interaction between two actors since it allows a formal, but still intuitive, definition of concepts like preference, rationality and trust between seeker and portal. Moreover, as will be seen in this chapter, it allows the creation of models at different levels of abstraction to describe theoretical, common practical as well as implemented portal solutions (section 5.2).

[29] A game-theoretic perspective on information search was introduced by us in (Eklund 2013b).

5.1 The search game

A *query* consists of a sequence of *words* posted as a single input to a portal's search engine. The query may be typed, or chosen from a list of predefined queries. The words may be expressed with both capital and lowercase letters. Since we will consider the searches not to be case sensitive,[30] we will by *query terms* refer to the lowercase versions of the words used in the queries. The *query length* is the number of terms in a query.

As part of the interaction between the seeker and portal, the query as well as information like the search time, session identifier, location of the seeker, the used device e.g. a mobile phone or a computer, and the portal page used to initiate the search are often registered by the portal, and by *context* we refer to this type of information.[31] By an *utterance* we denote a query and its, possibly empty, context.

When a seeker has posted her query, possibly with additional context information, to the portal she expects some "feedback" from the portal and an *answer* is a hyperlink to a *portal page*, i.e. a web page defined to be a part of the portal, considered by the portal to contain relevant information given an utterance. In the case of many answers these are presented as an *answer list* and one, called a *click*, may be chosen by the seeker for review. Each click has a *rank* corresponding to its position in the answer list. We say that a query is of a certain (average) rank whenever the clicks associated with the query has the (average) rank. A *session* contains the queries a portal user posts during a visit to the web site, their contexts, answers and clicks. Hence, the *search log* contains part of the session information for a given period of time. Each transaction in the search log is called a *search round*, and if the answer list is empty or the seeker did not choose any answer the round is called *incomplete*, otherwise it is a *complete round*.

Utilising situation-anchored game theory, we may say that at the beginning of a session the seeker is in a given situation s posting a query φ. Thereby the seeker is the actor who makes the first move by her utterance $\hat{\varphi}$ corresponding to the query φ and the associated query context $c(\varphi)$. When the seeker has made her move she is in a situation t awaiting certain types of actions from the

[30] Obviously the use of upper and lower case letters makes a difference, but due to the inconsistent use in our sources for analysis of the search logs, it will only be considered in sections 6.6.2 and 10.2.1 on studies of acronyms and capitalisations as means to indicate medical concepts and locations.

[31] It is not clear if the context which is registered by the portal is to be considered to have been provided by the seeker, even in cases when the seeker may not be aware of it and its consequences, see discussion in sections 4.6.2 and 9.2. However, in this work we treat the context as explicitly provided by the seeker as in the case of a query. Hence, the context may be seen as a query addition.

portal. Following the utterance, the second actor, the portal, makes its move σ, called the *answer*, by first, trying to *interpret* σ_i the situation t, possibly making use of the utterance and context $\hat{\varphi}$, *deciding* σ_d the appropriate answers, and finally *presenting* σ_p them to the seeker. The search round ends with the seeker in a new situation v, either being happy with a chosen click among the presented answers or initiating a new round, possibly by refining the previous query or leaving the session without choosing any answer. The sequence of utterance and answer can be depicted as in figure 11 with ϵ and δ reflecting the seeker's and portal's satisfaction with the outcome, respectively.[32]

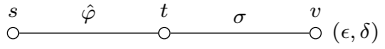


Figure 11: Schematic view of a search round.

The seeker's move(s) can be said to make up the *search strategy*, and the portal's move(s) make up the *answer strategy*.

Before we continue, let us walk through an example based on the following hypothetical situation:

> An evening in August at her summer house in Norrtälje, Julia suffers from a stiff neck and fever. Using her smartphone she searches for information at vardguiden.se by posting the query *stel i nacken feber* 'stiff in the neck fever' to seek advice.

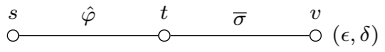
By the description, the query φ is the expression *stel i nacken feber*, and we assume the portal registers the context $c(\varphi)$ possibly represented by the sequence (evening, August, Norrtälje, mobile). Julia's situation s, the *pre-query situation*, is "information seeking", and following the utterance her situation t has changed to "advice expectation", figure 12. Hence, Julia's move consists of the query φ and the context $c(\varphi)$, which is also the information available to the portal. Thereby, we assume the portal always has access to the seeker's move when making its move.[33] The portal does not know Julia's expectations, the *post-query situation* t, but has to interpret her situation based on her move, and try to estimate $\overline{\sigma}_i$ her needs.[34] This may then help the portal to decide $\overline{\sigma}_d$ the answers to provide, and finally to present $\overline{\sigma}_p$ them. For instance, in this

[32] This discussion was limited to sessions of only one query and answer, but can be extended to more complex queries and interactions. Moreover, the actual process of the portal providing the answer and the seeker's interpretation and resulting situation may also be included in the model.

[33] In a game-theoretic terminology this implies that the game is dynamic and sequential, and with *perfect*, but not *complete*, *information* available to the portal. See also section 4.6.1 on information and situations.

[34] As will be discussed in section 5.2 on modelling seekers' behaviour, we will use the no-

example the portal may have no difficulty interpreting *feber* 'fever' as a symptom, but the rest of the utterance may prove to be slightly more problematic. Hence, the interpretation of *feber* as a symptom is an approximation of the medical term of (suffering from) fever. The portal did not pay any further attention to the query context other than making sure the decision on answers were constrained to Norrtälje, if these refer to care providers.



$$
\begin{aligned}
s &= \text{information seeking} \\
\varphi &= \text{stiff in the neck fever} \\
c(\varphi) &= \text{(evening, August, Norrtälje, mobile)} \\
t &= \text{advice expectation} \\
v &= \text{increased disease knowledge} \\
\epsilon &= \text{1-12/20} \\
\delta &= \epsilon
\end{aligned}
$$

Figure 12: Example of a search round.

With the utterance as input, the decision is made by the portal to provide answers related to diseases with given symptoms and present these as an ordered list of twenty alternatives. In other words, the decision trying to satisfy the needs in situation t was to assume that symptoms are provided when information on diseases with these symptoms is desired. Thereby the portal has made its move $\overline{\sigma}$ consisting of the interpretation, decision and presentation of answers, and since none of the pre- and post-situations were known, the move had to be based on approximate reasoning utilising available information such as the query and its context. Julia will then review the alternatives and choose the 12th answer of 20 related to a tick-borne disease called TBE, corresponding to an *answer situation* v with a payoff ϵ reflecting her satisfaction in comparison to other of the possible strategies she could have chosen when deciding which query to post and answer to pick. The portal was able to provide answers not ending up with an incomplete round, but noting the rank of Julia's click, its satisfaction δ could possibly have been better, since the portal's "score" at the end of this game is based on the rank being as low as possible.

---

tation $\bar{\cdot}$ (overline) to denote an approximation of the concept $\cdot$. For instance, $\overline{\sigma}_i$ indicates an attempt to, as well as possible, mimic the behaviour of the theoretically best possible interpretation σ_i.

It is important to emphasise that the seeker situations are not captured by the context. Only the query and some context aspects like seeker location, used device and time of search may be made available to the portal. Moreover, the only answer situation information the portal obtains from the seeker, and vice versa, is at best the chosen answer, or that it was the m'th answer among n provided. Hence, the model described in figure 13 may be seen as an approximation of the theoretical one in figure 11, and model approximations will be discussed in section 5.2.

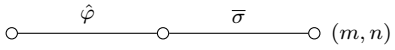


Figure 13: Schematic view of a search round as captured by a search log.

To summarise, we define a *search game* as a situation-anchored sequential dynamic game with bounded rationality of two actors, called the *seeker* and the *portal*. In this game the seeker always makes the first move, i.e. the *utterance* to reflect her being in a corresponding situation. The portal's moves are called *answers*, ending up with the seeker in a situation with a provided list of answers to act on. We also assume the actors to have perfect, but not complete, information of the other's moves.[35]

If we try to rephrase the presentation above in terms of the discussion in section 3.2 on communication expressed in terms of writers and readers, the portal is the reader and the seeker is the writer, whenever the latter posts a query. However, by the use of the context we might view the meaning available to the reader to consist of both a part found in the writer in addition to the one found in the query. Hence, our game-theoretic perspective on information may be seen as an attempt to, mathematically, describe a communication act where consideration is taken by the reader to both the meaning found in the writer and the text. Thereby, being interested in both the expressive and poetic functions of the communication, but from a portal provider perspective also the conative, referential, phatic and metalinguistic functions are of interest.

5.2 Describing and predicting seekers' behaviours

As discussed in chapter 3, two of the main aims of our work are to be able to *describe* and *predict* information seekers' *behaviour* based on search log information. To exemplify this, we begin in section 5.2.1 by presenting an

[35] Formally a search game requires an underlying bounded rationality, but we will use the notion of search games in discussions related to analysed search logs even when the preference relation is unknown and thereby not proven to adhere to the requirements of rationality.

example of the importance of a health portal to be able to predict the seeker's needs in a mobile context.[36]

In section 5.2.2, we continue the presentation already alluded to in section 5.1 on the challenges for the portal, but also the seeker, to understand and react on the seeker needs in a fruitful way, and how this leads to three levels of search games which we call Utopia, Template and Instance.[37] The first level of models correspond to the ones describing the theoretical foundations whenever situations are known, and Template is a template for models describing examples such as the one in section 5.1, where approximations are used based on available information. Finally, the Instances represent specific examples of realisations of the Template.

5.2.1 An example

When an information seeker has submitted a query (φ, figure 14), e.g. *stroke*, to the portal 1177.se, she could be expecting information on, for instance, *aetiology*,[38] *prevention*, *treatment*, general disease *information*, patient *stories* and information on specific *care givers* (post-query situation t_i). The reason for searching may be that she, or a relative, has *symptoms* of a disease, has (been diagnosed with) a *disease* or is generally *curious* about different aspects of a disease (pre-query situation s_i). Hence, the pre-query situation reflects the *reasons* for searching and the post-query situation captures the *expectations*.

However, the only information accessible to the portal is the utterance ($\hat{\varphi}$), and possibly historical information on these in relation to answers and clicks, trying to capture the answer situations (v_i) of the outcome and satisfaction for the seeker. For instance, in the case of searching 1177.se (Västra Götaland) with the query *stroke*, the clicks can be divided into ones related to aetiology (6.7%),[39] prevention (23.2%), treatment and rehabilitation (5.8%), general information and facts (11.0%), patient stories (43.7%) and information on specific care givers (13.1%), table 5.1. Hence, we hypothesise the answer situation v_i of the seeker to be possible to approximate with the chosen type of click, with proportions of search rounds corresponding to the probability of the seeker being satisfied. That is, if the portal only bases its move, or answers, on the seeker utterance and past clicks, then among the list of answers the seeker

[36] This example is also discussed in section 6.2.1.

[37] These discussions are independent of the notion of search games, and could have been made part of the primer on game theory. However, to keep the presentation accessible, we decided to include it in this part of the thesis.

[38] *Aetiology* refers to causes, or manner of causation, of a disease or condition.

[39] The numbers are percentage of search rounds, where the clicks have manually assigned types of answers based on the available click information, i.e. portal hyperlinks.

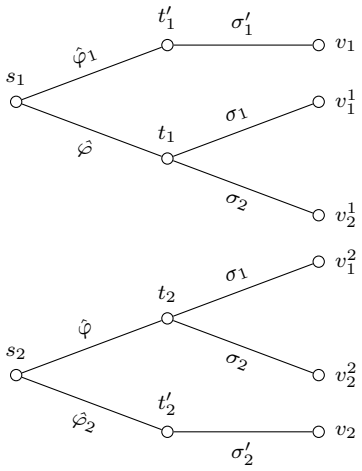


Figure 14: Game-theoretic view of information search.

will, for instance, in 5.2% of the cases select a rehabilitation-related answer on fatigue. If we assume the portal was to present only one type of answers, based on its prediction of the seeker's post-query situation, the risk would be that in 94.8% of the cases a seeker interested in fatigue in relation to stroke would possibly not be satisfied with the provided answers. This is supported by the considered page to be the only one addressing this rather common problem, and neither the title nor the gist mention the word *trötthet* 'fatigue', or one with similar meaning. Moreover, from a portal perspective a safe choice would be to always present patient stories, since they make up 43.7% of the preferred answers. Consequently, query and click information do not reveal which states, or situations (s_i, t_i), the seeker was in when the interaction took place. In other words, if the portal treats the proportion of search rounds for a given query as the only indicator of how to answer, the best odds, still less than random, of a satisfied seeker is obtained when she is interested in patient stories.[40] Thereby, the portal might tend to present a longer list of several types of answers to hopefully satisfy the seeker.

Receiving a long list of answers is not feasible in a mobile setting where the ability for interaction by browsing among different suggestions is limited. For instance, people with an interest in fatigue in relation to stroke (5.2%) are at risk of not being able to, among very few suggestions, find the relevant answer, and, making up a minor part of the seekers in the stroke context, also are at risk of not being accounted for by portal improvements focusing on the

[40]The figures do not take into account the rank of the answers. Hence, this reflects the assumption that the chosen suggestion was always the most preferred one by the portal.

Click	Search rounds	Rank	Answers	Answer type		
Magnus and Lillemor's story	80 (24.5%)	1.0	217.4	Inquiry	Disease	Group (patient)
SU Stroke Forum	9 (2.8%)	2.0	217.0	Facility	Location	SU
SKAS Neurology	2 (0.6%)	2.0	127.0	Facility	Location	SKAS
Facts I (slaganfall)	27 (8.3%)	2.4	111.1	Inquiry	Advice	Disease
SU Geriatrics	13 (4.0%)	3.0	216.0	Facility	Location	SU
Johanna's story	16 (4.9%)	4.0	214.8	Inquiry	Disease	Group (patient)
Marie's story	41 (12.5%)	3.7	219.5	Inquiry	Disease	Group (patient)
Marie's story	6 (1.8%)	10.0	110.0	Inquiry	Advice	Group (patient)
KS Geriatrics	9 (2.8%)	5.0	214.0	Facility	Location	KS
Stress (aetiology)	22 (6.7%)	5.3	215.9	Inquiry	News	
Problems (rehabilitation)	17 (5.2%)	6.9	211.4	Inquiry	Disease	Location (regional)
Unknown facility	1 (0.3%)	7.0	107.0	Facility	Location	Unknown
SKAS Stroke Centre	9 (2.8%)	9.0	217.0	Facility	Location	SKAS
Facts II	9 (2.8%)	11.0	217.0	Inquiry	Disease	Disease
Prevention	2 (0.6%)	14.0	102.0	Inquiry	Advice	Location (regional)
Prevention	74 (22.6%)	2.9	216.7	Inquiry	Disease	Location (regional)
Dental care (rehabilitation)	2 (0.6%)	15.0	102.0	Inquiry	Advice	Location (regional)
	Total: 327	Avg: 3.7	Avg: 204.2			

Table 5.1: Statistics for the query *stroke* posted in Västra Götaland (1177.se) from June 2010 to September 2011.

major groups. Hence, portal developers in the era of mobility are faced with the challenge of improving ranking of clicks, where in the case of people searching with the query *stroke* answers with an average rank of 3.7 were chosen. For our group of seekers, the average rank is 6.9 for the answer related to fatigue.

One approach to address the challenge of providing few but good answers is to try to improve the portal's ability to interpret the utterances, i.e. to better decide adequate $\overline{\sigma}_i$. Another path is to "force" the seekers to improve their expressions, thereby utilising the payoffs as a means to change the seekers' behaviour. The first approach may seem like the obvious path with emphasis on improving the portal interpreters. For instance, a basic approach to provide better answers is for the portal to try to predict the state t_i (e.g. interest in stroke rehabilitation) of the seeker by letting her answer which of the, in this setting, six classes of information (e.g. treatment, prevention etc) she is interested in. However, if the portal would try to be even more focused in its choice of suggestions, the seeker would have to choose from a large number of alternatives, in the case of stroke more than 20 different answers, and she would go from searching to browsing for relevant information. In our example, where rehabilitation is the least chosen interest (5.8%), there is a risk that this type of suggestion would end up at the end of the proposed answers. In the second case, we could try to "force" the information seeker to better define her interests by passing more "detailed" queries (φ_1). For instance, if the seeker would have used the query *stroke trötthet* 'stroke fatigue', there was only one answer which would match the query. Hence, if the portal is able to change the behaviour of the seeker towards certain types of utterances, then the risk of misinterpretation by the portal would decrease and the seeker would obtain more adequate suggestions.

If we now try to describe these characteristics of the portal graphically it may look as in figure 15. Based on the assumption that the portal's move only considers the actual utterance, in this case with an empty context, all needs will originate in the same post-query situation for stroke. Consequently, each query is associated with a unique post-query situation $\overline{t}$. Considering the pre-query situation $\overline{s}$, we assume it to reflect the overall aim of the portal 1177.se to provide information on health and disease matters of public interest. The answer situations describe the actual clicks, their average rank and the proportion of queries ending up with choosing the click, which in combination with the average rank can be seen as a measure of satisfaction. That is, when the seeker posted *stroke*, in 5.2% of the cases she chose an answer related to fatigue with an average rank of 6.9. In other words, if we assume she was interested in fatigue in relation to stroke, there is a high probability she will have to read several answer gists, and if the portal promotes more common answers this will most probably be the case in the future. Hence, the portal's preference

relation is solely based on the indexing of the vocabulary of the portal and its set of portal pages. However, the figure clearly shows that the answer related to stroke and fatigue is the obvious one in the case of a the query *stroke trötthet* 'stroke fatigue'. Hence, if the portal was able to "force" the seeker to detail her query, or by other means, find out the interest to be related to fatigue it would be able to better support the seeker.

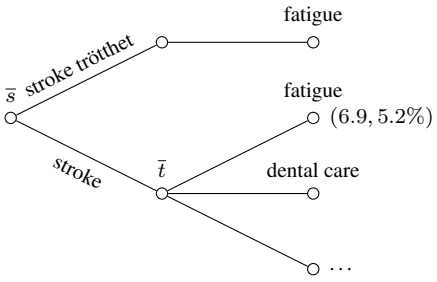


Figure 15: Schematic view of the query *stroke* as part of a search game related to fatigue.

If we analyse the clicks in detail, we have that the one related to stroke and fatigue is one of two related to rehabilitation, hence we may assume these answers to be related to a post-query interest in rehabilitation, figure 16. Thereby, we have that a portal able to interpret the utterance to be related to rehabilitation would be in a better position to satisfy the seeker. Moreover, we notice that the considered answer is obtained whenever the seeker searched with a query comprising a *Disease or Syndrome* term (stroke) and a *Sign or Symptom* (fatigue), hence we may hypothesise that for some diseases this type of queries indicate an interest in disease rehabilitation, or *Therapeutic or Preventive Procedure*. This can now be described as a potential search game, figure 17, worth further investigation if it reflects the way seekers behave, figure 18, and potential implementation.

By this example we have tried to show why it is important to be able to describe and predict information seeker behaviour, especially to be able to describe it at different levels of approximation depending on if focus is on specific types of solutions or more fundamental questions.

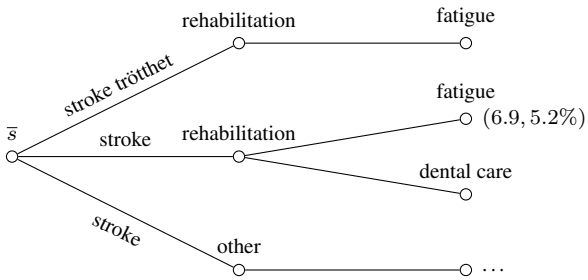


Figure 16: Schematic view of the query *stroke* as part of a search game related to fatigue, with answer type taken into consideration.

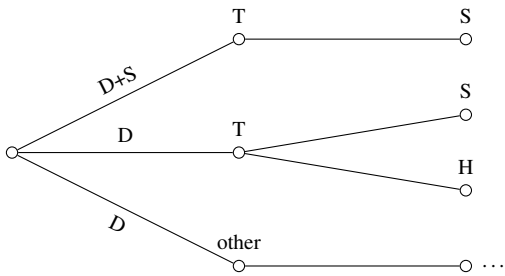


Figure 17: Schematic view of a hypothetical search game for rehabilitation related queries, where the abbreviations denote *(D)isease or Syndrome*, *(T)herapeutic or Preventive Procedure*, *(S)ign or Symptom* and *(H)ealth Care Activity*.

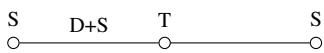


Figure 18: Schematic view of seeker behaviour related to interest in rehabilitation, where the abbreviations denote *(D)isease or Syndrome*, *(T)herapeutic or Preventive Procedure* and *(S)ign or Symptom*.

5.2.2 Different levels of models

In figure 11 (section 5.1) we saw how the interaction between an information seeker and a portal can be theoretically described, and above we discussed the importance and impact of the utterances and the portal interpretations to satisfy the needs of the seeker. However, as emphasised, the actual seeker situations and portal behaviour are unknown to the actors, and at best intelligent estimates may be obtained to help the interaction. Hence, there will always be

a gap between the "utopic" understanding σ of the seeker's post-query situation, and expectations, t and the realised interpreter $\overline{\sigma}$ utilising the captured context $\overline{t}$ of the situation, reflecting the incomplete information of the search game. Obviously, there are many different approaches to try to minimise this gap, but all these "instances" will share some basic assumptions like bounded rationality of a (theoretical) "template" for feasible solutions.

In this section we address the concepts of template, instances and their relations to a utopic model, with the former as levels of approximation, in a theoretical sense, and as levels of implementation in a practical sense of a portal which would always provide the best possible answers to satisfy the seeker's needs. In chapter 6 we elaborate on how these levels of models can be derived from existing interaction information, and in part II we discuss existing and future portal solutions based on an analysis of two official Swedish health information portals.

We start by recapitulating the model of the best possible world, called the *Utopia* model, corresponding to the situation when every actor explicitly knows the other actor's intentions and behaviour (cf section 5.1). Then we describe a model, called the *Template*, for approximations of the Utopia and how different behaviours are described and managed in the Template to derive further approximation models called *Instances* to reflect their shared features with the Template, but also the intention to describe existing, or future, portal solutions.

5.2.2.1 The Utopia model

The Utopia model, figure 19, describes the setting where the seeker knows how the portal would behave if it knew the seeker's query φ and pre- and post-query situations (s and t), and the portal knows the different answer situations v for different behaviours (with ϵ and δ reflecting the seeker's and portal's satisfaction with the outcome). Moreover, both actors are fully aware of both actors' preferences. This model with perfect and complete information is obviously, as implied by its name, impossible to reach in practice. But if it was achievable, any seeker query would result in only one, and the best possible, answer presented to the seeker. Consequently, the best we can achieve are attempts to estimate, or approximate, the different parameters of the Utopia based on the available information. Still, the intentions will reoccur in the ones trying to approximate the Utopia search game.

For instance, in the best of all worlds the seeker in the example in section 5.2.1 would express herself in such a way that if she was interested in information on fatigue in relation to stroke rehabilitation, cf the use of the query *stroke*

trötthet 'stroke fatigue' in figure 15, the portal would "know" this and provide only one answer to the portal page on the topic.

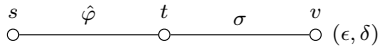


Figure 19: Basics of the Utopia model.

5.2.2.2 The Template

The aim of the Template is to identify the, theoretically, available information and its place in the context of a search model. Thereby any practical realisation of an information search solution can be seen as an instance of the Template. Hence, as the name implies, this model can be seen as a template for other more instantiated models describing different practical solutions. For instance, the models in figures 16–18 can be seen as attempts at different levels of approximation to describe properties of a potential implementation of a portal solution for certain types of queries.

To begin with, the only information the portal can obtain from the seeker is the query and its context. The latter can be seen as additional query terms to filter, or guide, the portal in its efforts. Thereby, as discussed in section 4.6.2, the concept of situation can be modelled by traditional game theory with bounded rationality with respect to query contexts. In addition to the query and the context, the log may contain information on provided answers and clicks, possibly as search sessions. Thereby, the search log is a transcript of the, theoretically, perfect information of the search game.

Figure 20 depicts these aspects of an interaction known at the end of a play, $\overline{v}$ represents the registered information about the answers and clicks and $\overline{t}$ is the (derivable) information on queries and contexts as captured by the search log. Hence, the knowledge which is possible to induce from the search log is made up by the two parameters $\overline{t}$ and $\overline{v}$, and is used when approximating the outcome of a fruitful play of the Utopia game. Thereby, the preference relation among answers for the portal is defined with respect to the query context $\hat{\varphi}$ and information in the estimates $\overline{t}$ and $\overline{v}$ of the post-query and answer situations, respectively. Consequently, the portal prefers answers, or moves, which are supported by the existing information on previous registered seeker portal behaviours for given queries and contexts.

The relation between the Utopia and Template can be summarised as in figure 21, where we have three approximations taking place.

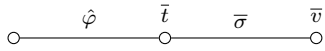


Figure 20: Basics of the Template model.

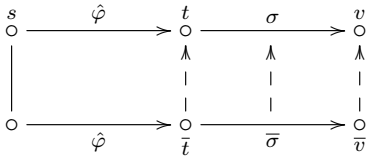


Figure 21: Relation between the Utopia and Template models, where the dashed arrows denote approximations.

5.2.2.3 The Instances

As described above, the Template model can be used as a template for the different realisations of a portal solution and its interactions with an information seeker, for instance, different ways the portal may make use of previously preferred answers by the seekers given their queries and contexts such as used device and location. In these cases the parameters $\overline{s}$, $\overline{t}$ and $\overline{\sigma}$ will be instances of the Utopia model following the constraints outlined by the Template.

Figure 15 describes another example with a simple index-based portal implementation where no care was taken regarding the expectations of the seekers, and answers were basically decided by a lookup of query terms in the index of the portal pages. Moreover, the instance tells us that the post-query situation is approximated by the query, hence disregarding information such as location, time and mobility. However, figure 17 describes a template for a family of instances where query terms are mapped to concepts and semantic types, for instance, *stroke* to *Disease or Syndrome*. This additional information will then be used as an approximation of the seeker's post-query situation to guide the portal towards certain types of answers. The last step is to present these answers as an ordered list, and register the adequate information, e.g. the number of answers and the position of the one clicked, in addition to the clicked link and query. All these considerations can be captured by a portal preference relation with bounded rationality with respect to the semantic information regarding the query.

Before continuing, it is worth taking a moment to discuss our presentation of instances in some detail. Firstly, our aim is not to describe exactly how a portal functions, but to highlight some important aspects in the context of the overall aims of this work. Hence, both the descriptions and the notations will

look rather sketchy in the eyes of a "formalist". Secondly, we will in most cases prefer to describe the instances in a narrative form to reduce the amount of formalism. Finally, we assume the reader to be able to translate our descriptions into formal ones, and implementations, if wanted. Hence, the presentation of an instance aims at conveying the ideas of portal realisations, and not its formal mathematical description or its implementation.

When talking about instances we intend models, possibly describing implemented solutions, sharing fundamental aspects described by the Template, which in turn is a theoretical approximation of the ultimate portal solution. This can be summarised as in figure 22, where the Instance approximates the Template, which in turn approximates Utopia.

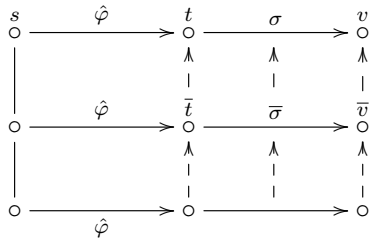


Figure 22: Relation between Utopia, Template and Instance models, where the dashed arrows denote approximation.

5.2.3 Search scenarios

The model in figure 18 does not only reflect a template for implementations, but also a potential *search scenario* where the seekers suffering from a disease by posting symptoms expect to obtain information on these in the context of their disease treatment (rehabilitation). Hence, the introduced framework and the levels of models also invite to descriptive and predictive analysis of different types of search behaviour, independent of the realisation of portal solutions to address them.

These scenarios may not only reveal features of how information seekers search for information, but also why. For instance, in our stroke example we used knowledge about the content of the clicks in addition to the submitted query to hypothesise the post-query situation, or the seeker expectations. Knowing the portal page to address problems in connection with stroke, and the seeker to post the query *stroke*, we may assume there is an implicit problem, or symptom, framing the post-query situation and being the focus of the pre-query situation of the interaction. Moreover, as depicted in figure 18, a general

hypothesis on queries and expectations may be stated for further theoretical discussions, experimental analysis or potential portal improvement.

5.2.4 Preference induction

The construction of the models in the preceding sections began with a search log and a basic statistical analysis of patterns of utterances, answers and clicks to be used to describe the possible paths in the models from pre-query to answer situations. Thereby, we might be able to use this approach to induce preference relations describing the "machinery" generating the outcomes found in the search logs, and as visualised in search games. However, the general task of inducing a preference relation from the information on plays, often called the *inverse problem* of game theory, is a non-trivial problem, see for instance (Parfionov and Zapatrin 2011), and out of scope of this thesis.

Even if the induction of preference relations is a difficult problem, an analysis as the one carried out in this chapter to describe and understand the interactions related to stroke reveals features of a potential preference relation. For instance, the implemented preference relation may have problems dealing with certain types of seeker interests as ones related to fatigue in the context of stroke rehabilitation. Moreover, the analysis highlighted a potential scenario of therapy interest expressed by queries made up of combined disease and symptom queries (figure 18). Consequently, describing a potential portal preference relation at an abstract level. This may then be used as a basis for further efforts to define and implement a preference relation being rational, and prompting a trustworthy portal behaviour. This approach is further exemplified in section 9.2.2.1 (see also section 6.2), where an analysis of context dependency is used as a basis for outlining a potential portal improvement shown to originate in a rational preference relation.

5.3 Related work

In the introduction to the thesis (chapter 3) we mentioned our work to be inspired by the ones of Parikh (2010) and Parfionov and Zapatrin (2011), and in this section we will briefly recapitulate some of the cornerstones and results of these efforts.

Parikh's work can be summarised as by the Nobel Laureate in Economics Eric S. Maskin as an "intriguing mixture of linguistics, computer science, game theory, and philosophy. It does much to illuminate an enduring mystery: how language acquires meaning". (Parikh 2010: back). In the introduction to the

book (2010: 1), the author states it as

> "In this book, I present a new account of meaning for natural language. The account has three levels. Most concretely, it offers a tool to derive and compute the meanings of all possible utterances, at least in principle. More generally, it provides a method to produce variant theories of meaning and to address the many problems and puzzles that beset its study. Most abstractly, it advances a way to think about meaning and language through the lens of a broad and powerful idea and image."

His framework is grounded in the theories of games and situations to capture meanings of any kind of utterances making up the communication between actors. The role of the first theory is to provide a setting for both the *microsemantics*, i.e. communication between and among individuals, and the *macrosemantics*, i.e. the "attributes of language that emerge in entire populations", in a similar way as game theory is used to study micro- and macroeconomics (Parikh 2010: 32–33). As a result of his approach, he provides the notion of *content* which joins the traditional ones of semantics and pragmatics, and is the *equilibrium* where the speaker's choice of utterance and the addressee's interpretation are in *balance* (2010: 26). Moreover, Parikh introduces four types of mappings, called *constraints*, where the first one, called the *syntactic constraint*, describes the syntax, or grammar, of the used language. The *conventional constraint* then maps the word of the language to properties and relations among the words which are independent of the context of the communication, and the *informational constraint* places these into a situation-specific environment. Finally, the *flow constraint* ties these constraints together as a system of building blocks to obtain the content, or *equilibrium semantics* of the communication act. For further details, we recommend the book "Language and Equilibrium" by Parikh (2010), especially the first chapter, which provides an introduction to background and challenges addressed by the author.

Parikh's framework is a hefty piece of work with a substantial amount of mathematical machinery to allow formal definitions and reasoning, and in theory a computable framework for implementation. As stated in the introduction to this thesis, Parikh's effort introduced us to the use of game theory as a tool for studies regarding communication providing a simple, yet powerful framework for topics of interest. However, the use of game theory to study communication is not new, or without critics (Allott 2006), but we would like to emphasise that our interest is not in equilibrium semantics as such, but in the use of game theory as language of modelling, which could attract theoreticians as well as practitioners, where both the information seekers and portals play equally important roles.

The other work which influenced us was a paper by Parfionov and Zapatrin (2011) where they present a macro-model of information retrieval based on game theory. The authors treat a search log as a transcript of so called *Nash equilibrium strategies* and address the inverse problem of finding the appropriate payoff functions, hence the rules of the game. Thereby, the paper focuses on the question of if, and how, it is possible to induce the behaviours adapted by information seekers and portals in their efforts to reach a state of mutual satisfaction. As in the case of Parikh's framework, it is not the authors' results as such that we use as a basis in this thesis, but the view of information retrieval as a game of two (communities) of actors, where the transcript, the search log, describes the play and how it may be used to obtain an understanding of the rules of the game and to predict, or simulate, future plays.

Game theory in the setting of communication has also been used by, for instance, Bodoff (2009) to describe and promote evolutionary forces to improve indexing and Jain and Parkes (2009) and Jain, Chen and Parkes (2009) to study the role of incentives in online answer forums.

5.4 Comments on the choice of a game-theoretic model

In this chapter we have tried to describe how information search can be described in terms of a game-theoretic view as an interaction between two actors. However, as alluded to in chapter 3 there are many different ways to achieve similar results. In our opinion three of the most appealing reasons for our choice of model is its simplicity at a conceptual level as an analogue of parlour games, its underlying solid mathematical machinery if needed, and how it makes it possible to relate the important notions of preference and trust. It also offers an opportunity, not covered in this thesis, to both, in theory, automatically induce preference relations from search log data and to carry out what-if analyses of changed seeker and portal behaviours to study existing and future portal solutions.

The choice of model is not only based on its theoretical assets and opportunities, but also the fact that the only available data to be used as a foundation of a model were search logs corresponding to transcripts of actions without any further information on the social or emotional aspects before, during or after a search session. Hence, we did not have an opportunity to address many of the cognitive features discussed in proposed models in, for instance, (Dinet, Chevalier and Tricot 2012). Moreover, by the nature of the interactions taking place in possibly stressful situations regarding sensitive matters and affected by both demographic, social and trust-related factors we considered models based on observational studies of ongoing interactions in real, or experimen-

tal, set-ups not to be feasible. Consequently, this also excludes models based on such data, even when those might better reflect the human behaviour of health information search.

6 From search logs to scenarios

In chapter 5 we presented a theoretical framework with search games for descriptive and predictive analysis of information search, especially interactions between health information seekers and public portals like vardguiden.se and 1177.se. In this chapter we outline how search logs can be used to induce search games, and highlight important aspects to consider in the process. We also discuss the role of annotation sources for analysis of induced search games.

This chapter can be seen as an *introduction* to our case study in part II as well as a *checklist* for those interested in pursuing similar studies in the future, based on the analysis of more than 15 million interactions from two evolving Swedish official portals over three years. The chapter addresses both the challenges and the opportunities, but most importantly, in our opinion, it *clarifies* the role of search logs and games as powerful and important tools for information providers to ensure improved support to portal users.

There may be many different reasons for why one would want to study search logs, but in this thesis we focus on, in our opinion, two fundamental topics of interest from both an information seeker and a provider perspective. The first is when and why do some queries end up with *no answers*, and what can a portal provider learn from studying these search games. The second topic of interest originates in the increased used of smartphones as a means to search for health information, requiring other user interactivity than traditional computers, and is when and why do some queries result in *many answers*, and which lessons there may be to learn from search games to avoid this. These topics cover different aspects addressing the ultimate aim of a portal which is able to understand the needs of an information seeker to present one and the best answer according to both the seeker and the portal. The topics are studied in chapters 9 (Queries without answers) and 10 (Queries with many answers) to try to describe *when* and *why* they occur, and to discuss findings of potential interest for information providers aiming at a trustworthy portal, summarised in chapter 11. To achieve this, we consider it important to divide the queries, and answers, in cohesive groups of entities with common properties and ap-

proaches to minimise potential negative impacts. For instance, considering the case with queries without answers there may be many different reasons for this reaching from basic typographic errors, ignorance of the language of medicine or the portal not "knowing" the used query terms. Obviously, each of these may benefit from different solutions, and we will re-use the work by James (1998) on *error analysis* to categorise problems occurring during second language learning to divide our queries into groups for further discussions. Similarly when dealing with queries with many answers we make use of work on *stylistics* (see for instance (Carter and Simpson 1989) for a brief historical review) to describe different types of queries and their properties in the light of the answers provided by a portal.

One can always argue that understanding when and why certain search behaviours occur is neither necessary, nor sufficient, to solve any problems, and different algorithms may even do the job without having to consider these aspects. This may be true, but we believe there are benefits from trying to express and understand new problems in the language of existing ones as by the use of error analysis and stylistics. The first one mainly tries to describe problems in learning a new language and the other one is used to describe and analyse stylistics aspects of, for instance, literature. Hence, in our efforts we combine the theory of economic models with language learning and literature analysis methods, thereby also highlighting the benefits of scientific cross-fertilisation to address today's challenges in the era of mobility and information search.

For portal developers aimed at providing the best possible information to seekers in distress trying to find out what they are suffering from and whether to seek care, it is important to continuously improve the portal search engines, and the *search log*, e.g. the query, its context and possibly the number of answers and the chosen click, is probably the most important source to utilise. Even though the search log can be seen as a transcript of a number of rounds of a search game, it does not describe the actual play, especially it does not directly answer questions like why the information seeker and portal behaved the way they did. Hence, our task becomes to try to highlight when and how search logs still may enlighten portal providers about the plays taking place with their portals as one of the actors.

In section 6.1 we discuss the type of information which may be captured in a search log of a health information portal like 1177 Vårdguiden. To exemplify the topics, we make use of the following two hypothetical interactions, based on actual search log data, between information seekers and the health portals vardguiden.se and 1177.se:

> Diana lives in Stenungsund in Västra Götaland County. She posts the query *barnmorskemottagning* 'antenatal clinic' to the portal vardguiden.se and the portal presents no answers to her query.
>
> Mary is at home not feeling well and posts the query *feber hosta* 'fever cough' describing her symptoms to 1177.se. The portal presents 134 answers and she finds the gist of the third interesting and clicks the link taking her to a portal page about self treatment of her problems.

Section 6.1 ends with a summary of properties we believe a search log should possess to facilitate both analysis and portal improvements. The presentation in section 6.2 begins with the search game aspects which are directly obtainable from the search logs, and then we show how others may be estimated based on these. In other words, not knowing the Instance, nor the Utopia, we discuss how existing search log data can be used to, partly, describe the actual behaviour of portals, but also how this may support analysis on future improvements and decisions for portal design. We end the chapter with a discussion on how the presented material and methods can be used to identify patterns of interactions reflecting different *scenarios* and how *preference* relations may be induced from the search logs.

With our induced search games at hand, in the rest of the chapter we discuss approaches to analyse our two fundamental topics (section 6.3), utilising error analysis to categorise the types of problems leading to no answers (section 6.5) and stylistics to characterise queries related to many answers (section 6.6). However, as alluded to in section 6.4, the notions of *context dependency* and *interpretability* of queries are at the core of the analysis, by the context providing a setting for both interpretation of queries and interest of seekers. In section 6.7, these aspects are used in the setting of search games to define a number of principles we believe a (health) information portal should adhere to, cf properties of search game templates to be realised by their instances.

6.1 The search log

The most important type of information found in a search log is the actual queries posted to the portal, and in some cases every query is captured. However, it can also be the case that information is kept only for the queries which resulted in answers, or ones where the information seeker made choices among the answers and clicked a hyperlink corresponding to an answer of interest. In addition to the query, the search log may contain information on when it was

posted and if aspects like location and interest of the seeker are known, and in the rest of this section we will look at these different types of information and how they are used to establish models of the interaction between the information seekers and portal.

Based on our analysis of search logs from 1177.se and vardguiden.se, we would like to stress the importance for any analyst to study the logs of interest in-depth, not only to address her questions but to understand which interactions were captured, how it was done and how this may have changed over time. It is also important to try to differentiate between the interactions originating in seekers' moves and those reflecting portal actions possibly not known to the seeker. For instance, in one of our studied logs not only the query and answer of interest, i.e. one search round, were captured, but the portal also registered all possible types of answers in a way which could easily be wrongly interpreted as several rounds having been played.

In our example, the search log entries of Diana's and Mary's searches could have looked as in table 6.1, and in the rest of this section we elaborate on the captured features in detail.

	Västra Götaland	**Västra Götaland County**
Session	B2091...	CA1B9...
Time	2011-01-26 08:39:40	2013-04-23 11:30:16
Query	feber hosta	barnmorskemottagning
Click	Sjukdomar-och-besvar/ Egenvardsguide/ Sjukvardsradgivningen/ ?CatId=28356&ChapId=28357	–
Rank	3	–
Answers	134	0

Table 6.1: Example of search log entries for Västra Götaland (1177.se) and Västra Götaland County (vardguiden.se).

6.1.1 Query

A query consists of a number of words, and depending on the purpose of the model they may be normalised by making them case insensitive. However, in a stylistic characterisation of queries (section 6.6.2) case sensitivity, i.e. the use of capital letters, is useful to identify terms referring to places or ones being acronyms of diseases or procedures. Still, the main reason for not considering case sensitivity (section 6.5.2) is our annotation resources not using this

property in a consistent way. In our example, the basis for continued analysis consists of the query *barnmorskemottagning* 'antenatal clinic' in Diana's case and *feber hosta* 'fever cough' in Mary's, none of which would need to be treated in a case sensitive manner.

Not only do our used annotation resources not consider case sensitivity, but also aspects like an inconsistent way of naming concepts makes language processing methods as *lemmatisation*, i.e. the process of grouping the different inflected forms of a word as a single item, and *stemming*, i.e. the process of reducing inflected words to their stem or root form, in our opinion, less useful without major efforts to normalise the resources which was not in the scope of this thesis.

As part of considering incomplete search rounds (section 9.1) it is noted that a query is the result of an information seeker either typing or clicking queries suggested by the portal, and it is often difficult to differentiate between these two types of queries in the search logs. Based on our analysis, the latter type often consists of longer and more complex *suggestions*, such as *barn- och ungdomsmedicin* 'child and adolescent medicine'. However, in some cases the phrases provided by the portal, and acted upon by the seeker as if they were queries, were *instructions*, as *t.ex. mottagning eller typ av vård* 'e.g. health unit or type of care'. Hence, a basic query length analysis is one approach to identify these moves of suggestions and instructions, and differentiate them from actual seeker moves.

The query makes up the seeker's move in the search game, but as will be seen, the search log may also capture additional information. Our analysis of search logs also reveals a substantial number of queries, in addition to proposals and instructions, most probably not originating in a seeker's typing, or at least not indicating so by the query terms often consisting of portal generated technical user interface messages. It could be that these are related to seekers choosing user interface menu options, but to allow a unified use of search logs for portal providers, these will have to be transformed into "traditional" query expressions. Hence, analysts of search logs should make sure to treat these types of information in a sensible way in game induction analysis.

To summarise, queries are the foundation for any search log analysis, but we would like to emphasise the importance of understanding the types of queries captured in the log, some of them being portal induced suggestions and instructions and to be treated accordingly. Moreover, if methods like stemming and lemmatisation of queries are used, consideration has to be taken that annotation sources may not be adapted to these approaches.

6.1.2 Context

In addition to the actual query, a search log may contain information related to *when* and *where* the query was submitted. It could also contain information on, for instance, the used device for the search, and if the seeker constrained it to, for instance, focus on health care units.

From a search game perspective, the context captures part of the pre- and post-query situations. However, it is important to note that some of this information may be registered without the seeker being aware of it, and possibly with undesired results. For instance, our analysis (section 9.2) indicates that information on the seeker's location is both logged and used by portals to constrain answers in a way that can lead to portal behaviour which may seem irrational to the seeker. Hence, some of the context may not reflect the intended situation information the seeker would want to convey to the portal.

To summarise, context information is a valuable means for a portal to gain understanding of a seeker's situation, but as our analysis shows, care has to be taken to both the use of the information and if it actually reflects the seeker's situation and needs.

6.1.2.1 Time

Depending on the technology used to capture search information, the log may contain a timestamp of when a query was posted. This type of information is mainly of interest to order a search round with respect to other rounds in the same or other sessions. In our examples, the log tells us that Mary posted her query at eight o'clock in the morning of the 26 January 2011, hence in the middle of the flu and cold season, and Diana's was posted on 23 April 2013 just before lunch time.

Time information provides an order of the moves in a search game, but our analysis shows that this type of information is sometimes unreliable, as in the case of session information below. For instance, the log provides examples where many different clicks in the same session have been carried out the same second. Still, time information is probably a more reliable measure of chronology than session identifiers. Hence, it is a feature of the logs used to induce an order of moves, or rounds, in a search game.

6.1.2.2 Location

The log may also contain information on where the interaction took place, or the geographic area of interest to the information seeker. As seen in our anal-

ysis (section 9.2.2) of the use of 1177.se and vardguiden.se, this type of constraint is commonly used, sometimes possibly unintentionally set by the user, and with substantial impact on the interaction and the answers provided by the portal. In our examples, the log does not provide any further information than that both queries were posted in Västra Götaland, and Diana's location was specified to Stenungsund, which is a town in this county. Still, if this feature is used to decide answers to present to the seekers, they will make up part of the context used by the portal to decide its move and depending on the intentions of the queries may, in the eyes of the seekers, result in irrational answers as in Diana's case with no answers for a query which may indicate an interest in finding an antenatal clinic.

6.1.2.3 Other

In addition to context information on time and location, search logs may contain information on if a mobile device was used and if the seeker wanted the answers to be constrained to certain domains such as health care providers or facts on diseases. The latter type of constraints is mainly inherited from the structure of the portal. For instance, it may be divided into different areas, or pages, covering facts, administration, care providers etc, and when a seeker posts a query from any of these the answers are constrained to the corresponding type. However, it is unclear if these restrictions are known to the seeker and to be considered part of an intentional pre- or post-query situation of the seeker.

6.1.3 Session

The series of interactions between one information seeker and a portal may be identified by a *session identifier* corresponding to a play of search rounds. However, our analysis (section 8.1.1) has shown that this type of information seems to have different types of technical problems, such as reoccurring identifiers, and not be reliable for identifying search rounds and plays. Hence, models often have to be induced based on a division of search log entries into *intervals* based on the time of search. For instance, to consider all queries posted in a given time interval to reflect a single search play, see section 8.1.1 for further discussion.

The problem with the inadequate session information is one of the major obstacles to induce accurate models describing the interaction between information seekers and a portal. Based on our analysis, it seems that a majority

of all interactions are on the form of one query and one click, hence plays of one round and of length one. However, the search logs seem to contain substantial noise in the sense of interaction duplication, inconsistent registration of queries with respect to their contexts, consequently major efforts based on the characteristics of the search logs are needed to obtain a clean log useful for induction of models. In our example, Mary's question and click is the only round logged to be part of session "B209...", but due to known problems with session identifiers, we do not actually know that this was her only round.

6.1.4 Answer

A search log may capture information on answers and clicks at different levels of detail. For instance, for the 1177.se search log we have access to the exact click made by the seeker, but in other logs we only know the number of answers. In our example, Mary's query resulted in 134 answers, which possibly could be expected for queries referring to common symptoms as fever and cough. Diana, on the other hand, did not receive any answers. However, if Mary's query would not have resulted in any answers, we would not have known this since in her case the portal only registered complete search rounds. Hence, it is important to be aware of this type of differences in the choice and use of search logs.

As will be seen by our analysis of queries with many answers (chapter 10), detailed click information is a potentially very useful source for understanding seeker behaviour and to be utilised in both descriptive and predictive search game analysis, hence a means to improve interactions and seeker satisfaction. However, currently this information is sparse and not organised to facilitate these opportunities, see section 10.3 on answer stylistics.

6.1.4.1 Click

As seen by the example of Mary, it may be the case that the log contains valuable information on the types of answers which are of interest to a seeker. In our example we have that the click was part of a section maintained by Sjukvårdsrådgivningen on diseases with emphasis on self-treatment (table 6.1, part 28356 in chapter 28357). Hence, she clicked on a hyperlink providing information on self-treatment for some problems, which may not be surprising when the posted query was on symptoms such as fever and cough. Similar information may also be obtained by notes on if the answers were provided as part of a health care unit search. The level of detail and usefulness of clicks

from an analysis perspective depend on the portal structure, e.g. how hyperlinks mimic the structure of the provided information. As seen by our analysis in chapter 10, this type of information is very useful to better understand the interests of the seekers, but also the portal's behaviour. Thereby, being an important feature to ensure improved interactions between seekers and portal considering not only past queries and contexts, but also the provided answers and preferred clicks. But as seen in the case of the search log entry of Mary, we had to manually dissect and interpret the hyperlink information, and still not being able to reveal the actual information provided by a portal page with a given chapter identifier.

6.1.4.2 Rank

Finally, the log tells us that Mary decided to follow the third link in the provided list of answers, hence in this case the portal seems to have done a good job at providing the most "relevant" answers at the top of the list. The rank may also be seen as a measure on the amount of time a seeker dwelled on the answers before making a choice, based on the assumption that a seeker considers the gists of the preceding answers before deciding which one to click.

6.1.5 Desirable properties of a search log

Based on lessons learnt by the research presented in this thesis, table 6.2 summarises what we consider to be desirable properties of a search log to facilitate similar, and improved, analysis as herein.

6.2 Game induction

In the previous section we described the type of data found in a search log, and as seen by the discussion it is often in a state where it is difficult to identify the actual activities taking place and a division into sessions. Still, in our opinion, this is mainly due to inadequate capture procedures and should be possible to address in cases where information providers are interested in utilising interaction information to improve their ability to provide support to information seekers. Since this thesis is mainly about the use of search log information, and not how to register seeker–portal interactions, we assume steps are taken to try to minimise these problems and their impact and turn our interest to how search log information can be used to induce search games summarising

Aspect	Property
Query	As typed, including errors and case
Query	Differentiation between seeker queries and ones provided by portal as suggestions etc
Context – time	Time for query and potential click
Context – location	Preferably coordinates (respecting data privacy act)
Context – device	If mobile device was used for session
Context – domain	If search took place from specific part of the portal
Session	Unique identifier grouping all interactions in a session
Answer – click	Parsable and interpretable hyperlink
Answer – count	Size of answer list, including size per page of answers
Answer – rank	Rank at given page

Table 6.2: Desirable properties of a search log.

and presenting adequate features to allow descriptive and predictive analysis of seeker and portal behaviours and interactions.

6.2.1 From search log to Instances

Given sessions of queries and their contexts, instances of the interactions can be induced in a straightforward manner to facilitate analysis. For instance, Mary's interaction with the portal 1177.se can be depicted as in figure 23. The model tells us that the seeker posted the query φ, and in addition the context $c(\varphi)$ was registered. Based on the captured information, the portal approximates the seeker's post-query situation by $\bar{t}$ and makes the interpretation $\overline{\sigma}_i$ that the seeker has an interest in fever and cough. Hence, the seeker made a move $\hat{\varphi}$, resulting in the portal contemplating moves trying to satisfy the seeker's needs based on its interpretation $\overline{\sigma}_i$ of the situation. The portal's move is then made up of the pages with corresponding index terms, assuming pages with given terms to be good answers to the needs of the seeker, and to present them based on number of past clicks. Hence, the three steps of interpretation, decision and presentation make up the realisation of an underlying portal preference relation with a basis in the query and with preference for moves, i.e. lists of answers, which contain the query terms and which are more often clicked. The answer situation $\overline{v}$ captured by the portal consists of the number of answers n, the rank of the clicked m-th one and its hyperlink. Thereby, we have turned a search log entry into a play in a game, where the description of the utilised post-query information and the portal's preference relation defines the

"rules" of the game, and we may use the model to predict the outcome in cases with other queries. Moreover, the answer situation provides valuable information on how well the preference relation cares for the seeker needs, and how well it has interpreted the seeker's "message" for the portal to act on.

φ	=	$feber\ \ hosta$
$c(\varphi)$	=	(Västra Götaland, 2011-01-26 08:39:40)
$\overline{t}$	=	($feber\ hosta$, Västra Götaland)
$\overline{\sigma}_i$	=	"feber" AND "hosta"
$\overline{\sigma}_d$	=	pages with index terms "feber" and "hosta"
$\overline{\sigma}_p$	=	order by number of past clicks
m	=	3
n	=	134
$click$	=	$Egenvardsguide/Sjukvardsradgivningen/?CatId$=28356$ChapId$=28357
$\overline{v}$	=	(φ, $c(\varphi)$, m, n, $click$)

Figure 23: Example of induced search game for a query and its context.

Given the example in figure 23 with a context comprising Västra Götaland in late January, the model describes a setting where the portal's move is based on utilising an index of terms and data on past clicks. In this case, for any query the portal will always be able to decide for any possible pairs of answers which one that is best, i.e. the answer page with most past clicks for the given query. Moreover, if page A is better than B and B is a better choice than C, then the first choice A must have at least as many clicks as C for the given query. Hence, the underlying preference relation of the model is complete and transitive, thereby rational with respect to its context.

Considering the used dataset for Västra Götaland, Mary was the only one who had posted this specific query, but in most cases there may be many different seekers posting the same query with the same context (excluding the time aspect), possibly clicking different answers, and we have to decide how this is to be modelled. To discuss this topic we make use of the statistics in table 5.1 for the query *stroke* posted in the interval June 2010 to September 2011 in Västra Götaland.

Each record in the table can be turned into a search game as in figure 23.

However, they all share the query, even though the context may have differed. The number of clicks indicates how often given the query and context the seeker, and portal, ended up in a given situation. Moreover, the average rank tells us how "well" the portal preference relation generally reflected the seeker needs. If we try to annotate the clicks in a similar way as the query *stroke* as being of the semantic type *Disease or Syndrome*, we have that the answers can be divided into ones related to *Disease or Syndrome*, *Spatial Concept* (i.e. location), *Health Care Activity* and *Intellectual Product* (i.e. news).

By the concept and semantic annotations, we can perceive a pattern of interaction depicted as in figure 24, and hypothesise the portal to implement an Instance where it is almost equally difficult to find the answer of interest independent of the topic of interest, but with a tendency of disease-related answers to be preferred by both portal and seekers. However, it is important to stress that the actual implementation might be based on a basic algorithm as the one in our example with Mary, but its outcome to actually mimic a more sophisticated preference relation.

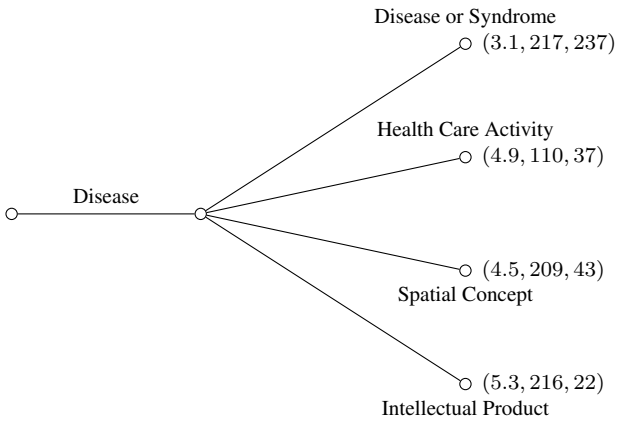


Figure 24: Interaction scenario based on analysis of stroke queries for Västra Götaland. The numbers in parentheses correspond to average rank, average number of answers and number of clicks.

As presented in section 5.2.1, and figure 24, it is possible, using search log information, to describe scenarios at different levels of detail trying to describe the implementation and common patterns of search.

At the instance level, the aim is to visualise the search log content in a way allowing a search game reasoning, i.e. introduce the notions of seeker and portal moves and estimates on how "well" the interaction turned out. Hence, the graph would look as in figure 25 for the stroke example. The next question is to find out how to connect the query and the answer information. In

other words, we need to define the estimated post-query situation $\overline{t}_{ij}$ resulting from the utterance $\hat{\varphi}$ originating in the unknown pre-query situation s_i. Given the captured search log information $\overline{v}_{ijkl}$ by the unknown portal move $\overline{\sigma}_{ijk}$, we can say that figure 25 describes part of the considered Instance, and depending on the amount of captured information and knowledge on seeker and portal behaviour we may further instantiate the Instance to reflect the seeker–portal interaction. For instance, we might as in figure 24 be able to divide the post-query situations into groups $\overline{t}_{ij}$, where j reflects the semantic type of the clicks and for each click obtain a specific answer situation (cf figure 15, section 5.2.1).

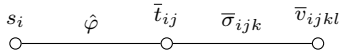


Figure 25: Schematic view of a search round.

6.2.2 From Instances to Template

As in the stroke example in section 5.2.1, it may be possible to find patterns among Instances as described in the previous section with certain types of queries leading to specific types of clicks (cf figure 17, section 5.2.1). These patterns, or potential *search scenarios*, can be found by utilising semantic annotation of both queries and clicks, and visualise the result in a similar way as for the instances. However, in this case the post-query estimate $\overline{t}_{ij}$ will represent, for instance, a given semantic type revealed by the answer annotation as indicative of certain seeker expectations. For instance, when a seeker clicks on stroke rehabilitation problems related to fatigue, we may assume the seeker's post-query situation to be related to expectations on rehabilitation, especially related to fatigue. Consequently, the answer situation estimate would also contain information describing it as seekers to be satisfied, as given by answer statistics. We might even as in figure 18 (section 5.2.1), be able to induce a hypothetical search game description of a search scenario solely based on semantic types and given context information.

6.2.3 From Template to Utopia

Finally, with a set of templates describing different scenarios we may try to establish a model of the Utopia estimated by the templates. For instance, that whenever a seeker posts a query containing disease and symptom terms, the portal should present the portal page describing the symptom in the setting of

the disease. Moreover, the portal page should focus on therapies to alleviate the symptoms.

6.3 Two types of seeker challenges

As discussed in the introduction to this chapter, our interest is to *characterise* the search rounds where the seeker and/or the portal run into problems such as no answers or too many answers. Thereby, our interest is not mainly to solve the problems but to describe the behaviours leading up to them, and in some cases this exercise also points in directions of potential solutions. Moreover, there are an infinite number of ways to describe challenging interactions, but in this work we lean on efforts in *error analysis* and *stylistics*. The aim of error analysis is to study the errors people make when using a language, especially those resulting from learning a new language. For instance, different types of spelling problems, lexical errors and pragmatic challenges (James 1998). Leech calls stylistics the child of linguistic and literary studies, based on the "idea that there is something special about literature" (2008: 2). One of the pioneers in the 1960s with an interest in the intersection of these two areas was Jakobson (1960) with attempts to define the concept of style and describe its characteristics, and for a brief historical introduction to the field see (Carter and Simpson 1989).

These two avenues to address no answers and too many answers are the focus of sections 6.5 (Queries without answers) and 6.6 (Queries with many answers), but we begin with a general aspect of importance to both these topics, the role of the context of queries and answers, which by the discussion above has been seen to be influential on the outcomes of search rounds.

6.4 Context dependency and interpretability

As presented in section 6.1.2, a query is often accompanied by context information, such as location and domain of interest, to possibly be taken into account when a portal tries to interpret the query and decide which answers to return. The context may be, more or less, decided by the seeker, possibly without knowing its implications. Our analysis of the search logs from 1177.se and vardguiden.se reveals a substantial number of queries resulting in no answers due to context constraints (section 9.2). For instance, queries about available emergency units in one part of Stockholm results in no answers at the same time as the query results in answers when posted in a neighbouring part of the city. Hence, the context has a substantial impact on the interpretation of queries

and on the portal's actions, see chapters 9 and 10 for further discussions.

Consequently, we need to be able to discuss this type of influences in our framework and call a query (term) *context dependent*, where context refers to aspects like location, time and search device, if the answers differ between at least two contexts (for a given dataset).[41] For instance, the query *akutmottagning* 'emergency unit' is *location-dependent* for Stockholm, since in Hägersten-Liljeholmen you may receive an answer but in Rinkeby-Kista this is not the case (section 9.2.1). Along a similar line of reasoning we call a query *time-* or *device-dependent* if the answers depend on time of search or device, respectively. A query (term) which is not context dependent for any pair of contexts is called *context independent* (for a given dataset). In a search game instance, a context independent query corresponds to the portal being able to decide its move only based on the query and no seeker situation information. In other words, the preferences and rationality of the portal considers only the way the seeker expressed herself as in the case with Mary in figure 23 where no attention is paid to the location information in the selection of answers.

A problem with the notion of context dependency from an analysis, and portal, perspective is that in implemented instances, such as the ones studied in this work, the search logs almost never contain any specific answer information, but only the number of answers, thereby making the notion of context dependency difficult to use. We call a query *weakly context dependent* (for a given dataset) if the numbers of answers differ between two contexts. Consequently, the query is called *weakly context independent* if the numbers of answers are equal for any pair of contexts. Hence, context dependency implies weak dependency, but not the other way around. However, as noted this is the strongest possible measure of dependency inducible from our search logs. The stroke example based on the information in table 5.1 reveals a weak context dependency by fluctuating number of answers, even though we have no information on query context or realised preference relations. As discussed in section 6.7 on properties of a trustworthy portal, context dependency is not a problem per se, but when the lists of answers differ over time it may seem to a seeker as if the portal is behaving in an irrational way. Hence, context dependency is not only related to adequate treatment of seeker situation information, but also to seekers' trust in a portal.

Based on the number of answers for a query, we call the query (term) *interpretable* with respect to a context if its answer list is not empty (for a given dataset). Otherwise it is *non-interpretable* with respect to the context. For in-

[41] Our notion of context dependency is not to be confused with the notion occurring in, for instance, grammar theory to denote grammatical constituents whose derivation or rewriting depends on its surrounding ones.

stance, in the Diana example above, the query is non-interpretable with respect to given location. The game-theoretic take on this is that a non-interpretable query leads to an incomplete round of the play, but may still seem rational to both players. A query which is interpretable with respect to any context is called *fully interpretable*, and if it is interpretable for only some contexts it is called *partially interpretable*. For instance, in our analysis the query *virus-prickar* 'virus spots' (table 9.15) is non-interpretable and *lever* 'liver' is fully interpretable for Stockholm County, both being valid medical expressions. This example also shows interpretability to be related to topics like vocabularies and spelling, thereby plays an important role in chapter 9 on origins of queries without answers.

As revealed in this work, the concepts of context dependency and interpretability are key aspects both in identifying challenging queries and in our discussions on the trustfulness of portals in the eyes of the seekers, where portals like 1177.se and vardguiden.se in our studies show behaviour leading to partially interpretable queries with respect to, for instance, both locations and time.

6.5 Queries without answers

As will be seen in our analysis of the search logs (section 9.2), we estimate more than 5% of the queries in the considered Stockholm regions and Swedish counties to be non-interpretable, thereby leading to incomplete search rounds. This may be a result of misspellings or the use of words not known to the portals. Hence, queries without answers are a problem which should be addressed by portal providers, especially in the light of the responsibility to provide adequate health related support to the public.

There are several different ways to study this type of problems, but we decided to view these as ones often found when trying to learn, or use, a language of less familiarity, e.g. the language of medicine used to describe symptoms, diseases and treatments on portal pages. Hence, we make use of notions from *error analysis* and begin with the concept of *vocabularies* to allow us to define terms as known or *unknown* to seekers or portals, but also to be able to denote terms as *misspelt*. With this as a basis we will, using the terminology of James (1998), categorise problematic query terms at several levels based on the way an information seeker *expresses* herself (substance), *lexico-grammatical patterns* (text) and the intentions of the seeker and *interpretations* of the portal (discourse).

6.5.1 Vocabularies and spelling

When discussing what is to be considered "misspellings" we need a reference to allow us to define a query to be *unknown* with respect to a set of words, called a *vocabulary*, if it is not a member of the set. Hence, the vocabulary can be seen as an enumeration of words, e.g. queries, possibly generated by a lexicon and a grammar. Thereby, a *misspelling* is an unknown word that by some defined modifications becomes known with respect to a given vocabulary.

In our setting of search games, the moves of the portal consist roughly of *interpreting* seeker moves, *deciding* on answers to provide and *presenting* them to the seeker. Hence, the first step is tightly connected to the queries submitted by the seeker, and these are, in theory, compared to an *index* of words considered to be known by the portal, and associated with given answers, or portal pages. Hence, the underlying index of a portal can be seen as making up its vocabulary, and a misspelling to the portal is a word transformable by a defined number of modifications into an indexed word. Otherwise it is considered to be unknown and, independent of the portal's interpreter, results in an incomplete search round.

If we turn our interest to the seeker, it becomes more difficult to define the vocabulary of the Swedish language, i.e. the potential "ultimate" set of words used by information seekers in the setting of health related questions. A simple approach is to define it to, in theory, correspond to the more than 125,000 words covered by the Svenska Akademiens Ordlista (2006), abbreviated SAOL, and an utterance to be unknown (with respect to the Swedish language) if it is not defined by SAOL. Depending on the intended use of the vocabulary, for instance as a computerised source for linguistic analysis, other sources like the Swedish Associative Thesaurus (SALDO) (Borin, Forsberg and Lönngren 2013) may be used as a basis to define a Swedish vocabulary. One can also use curated compilations of previously posted queries as a definition of the used language, or *seeker vocabulary*, assuming a large enough dataset. This approach is taken in our analysis in this thesis, where we have tried to manually exclude queries seemingly generated and registered by the portal, and not by the seeker as an explicit query or by choosing a portal suggestion.

Another problem is that there is nothing that prohibits the seeker vocabulary or portal index from containing words that are unknown with respect to the Swedish language. For instance, the query *Chrons* will be seen to be a member of the index of the Sweden dataset (section 9.3.1) even though the correct spelling is *Crohns* 'Crohn's'. Moreover, *Crohns* is not found in SAOL, thereby it should not be considered part of the language. Hence we need to extend the Swedish vocabulary with words found in, or inducible from, a standard

medical Swedish terminology like *Medicinsk Terminologi* by Lindskog et al. (2008) to obtain a better coverage of Swedish terminology possibly used by information seekers. As discussed in section 8.1.2 on used resources, the role of a limited Swedish medical vocabulary may be assigned to databases such as the Swedish Snomed CT (Socialstyrelsen 2013) and MeSH (NLM 2013b). To complicate things further, the queries *lpk* and *inflamation* will also, according to the Sweden dataset (section 9.3.2), both be non-interpretable, but in general we might consider the first word to be nonsense and the second one to be a misspelling of the known Swedish word *inflammation*. However, the term *lpk* is an abbreviation of the medical term *leukocytpartikelkoncentration* 'leucocyte particle concentration', thereby referencing a known word using an acronym.

To summarise, both a theoretical definition and a practical choice of vocabulary is a non-trivial challenge, but from a perspective of a seeker's trust in a portal, the portal should be able to interpret queries not obviously, to the seeker, being misspellings and making sense in the setting of a health portal.

Before continuing we need to elaborate on our choice of seeker and portal vocabularies used in the analyses in part II, to present how this type of challenges can be addressed. As the portal vocabulary we choose the set of queries found in a search log resulting in non-empty answer lists, keeping in mind that the portal may contain portal pages, or answers, never clicked. When deciding on the seeker vocabulary, the used terms in queries, a simple approach is to let it consist of all queries posted by seekers, trying to exclude the strings logged as queries but resulting from unclear portal actions and often containing non-alphanumeric symbols. Still the problem remains that this vocabulary will also contain queries which are misspellings with respect to any of the choices of Swedish vocabulary above. Moreover, by the discussion there is to our knowledge no automatic way to identify these misspellings since either, as in the case of SAOL, the vocabulary is not available in a useful format, or as in the case of SALDO, it does not contain the medical terms used by information seekers. The latter problem is also shared with medical sources like Swedish Snomed CT and MeSH. Hence, there is no obvious choice for a Swedish vocabulary for health information search. An approach, only briefly used in our work, is to combine any of the sources mentioned with methods trying to identify misspellings by comparing if and how query terms may be transformed into ones found in established vocabularies by a given set of modifications. However, this task has not been pursued in this thesis, and thereby we let the seeker vocabulary of a dataset consist of all queries, excluding the ones which seem to be portal generated. Hence, when referring to a misspelling we mean a query which is part of the vocabulary, but in our opinion should be considered to be a deviation. Hence, by the Swedish vocabulary we mean our subjective view on the words making up the Swedish language.

Following James (1998: 78), we define a linguistic "error" as "being an instance of language that is unintentionally deviant and is not self-corrigible by its author", and a "mistake" to be "either intentionally or unintentionally deviant and self-corrigible". However, now we run into the challenge of defining the notion of "deviant", and in this work we define a word to be *deviant* if it is not found in, or inducible from, a vocabulary of the ones discussed above.

James (1998: 83) further elaborates in his work on deviances and ends up with the following division, which we will make use of:

Slips are deviances which "can quickly be detected and self-corrected by the author unaided". For instance, typing errors such as *sjukhua* instead of *sjukhus* 'hospital'.

Mistakes "can only be corrected by their agent if their deviance is pointed out". In case an indication of the existence of a deviance is sufficient for the author to detect and correct it, it is called a *first-order mistake*, otherwise *second-order*. For instance, in the Sweden dataset, in 3.6% of all queries in which the term *inflammation* is included, it is misspelt as *inflamation* (section 9.3.2).

Errors "cannot be self-corrected until further relevant [...] input [...] has been provided", that is, they "require further learning to take place before they can be self-corrected". For instance, 12.7% of the queries in the Sweden dataset referring to the term *borrelia* are misspelt as *borelia* (section 9.3.2), thereby possibly indicating a second-order mistake, or even an error implicating no ability to correct the misspelling even if told it is incorrect.

Solecisms are "breaches of the rules of correctness as laid down by purists and usually taught in schools".[42]

As will be seen in our analysis of queries without answers in chapter 9, at least the first three types of deviances are found when studying the search logs, hence they may jeopardise the collaboration of the seeker and portal. Thereby, they are of interest to the portal to manage in a sensible way to satisfy the seeker's needs as well as possible, and maintain her trust by a rational behaviour in her eyes. It is also worth mentioning that in communication among humans, errors like the ones above often do not jeopardise the interaction in such a way that it will break down. However, the portal does not have the capacity of a human to interpret errors, but not knowing the capacities of the

[42] This type of deviation is not studied in this work.

portal, the seeker may expect more than achievable and the portal has to make sure that events at risk of leading to distrust are managed.[43]

As discussed by James (1998), there are many different ways to classify and describe language deviances, depending on the purpose of the task. In this work we will only make use of some of these to try to define and categorise the deviances found in our datasets, how they may be identified and their impact. With the search logs at hand, one obvious categorisation is to try to describe the type of "modifications", with respect to the Swedish vocabulary, that seem to have taken place. James (1998: 106–113) divides these into a *target modification taxonomy* of omissions, overinclusions, misselections, misorderings and blends:

Omissions are parts of words or phrases intentionally or unintentionally left out, resulting in invalid expressions. For instance, a seeker typing the query *åderråck* instead of *åderbråck* 'varicose veins'.

Overinclusions result from, for instance, inclusion of letters which should be dropped in compounds, e.g. *läkaremottagning* instead of *läkarmottagning* 'surgery'.

Misselections occur when writing, for instance, *vårdscentral* instead of *vårdcentral* 'health centre', that is the seeker selects the wrong form of a word in a compound.

Misorderings result from seekers placing (parts of) words in the wrong order.[44]

Blends refers to when e.g. synonyms are combined in an uncommon way in a compound, as in the case of the query *husdoktor* instead of *husläkare* 'general practitioner', where *doktor* and *läkare* are synonyms and *husläkare* is the established term.

The types of modifications can be seen as ways to describe the four types of deviations presented above, and areas to be addressed by the portal, and we are now ready to try to classify these according to their level of occurrence, that is if the deviation may be found at the *substance*, *text* or *discourse* level (James 1998: 129–172). According to James (1998: 130), this simplified classification

[43] It is also important to note that users of Google search engine are used to rather good ability of the search engine to interpret even substantially misspelt words, based on access to massive logs of user interactions as a basis for "correction" of misspellings. This is not an option for portals like the ones studied in this work.

[44] Since word order is irrelevant to the considered portals and we have not been able to find any words with misordered constituents this type of modification will not be considered.

reflects "standard views of linguistics" with the *substance–medium*, *text–usage* and *discourse–use* pairs. However, as will be seen by the presentation, and the types of deviations assigned by James to the different classifications, the rationale behind the division might not be completely clear, and in some cases the category names will be questionable. Still we consider James' terminology to provide a classification which fits our purposes. We will view the substance level as the actual expression, or *query*, the text level to account for the *lexico-grammatical patterns*, or ultimately if a word is known in the language, and finally the discourse captures the use of a query as part of an *interaction* with underlying aims and intentions.

Before continuing, it is worth elaborating on our use of James' categorisation, and why we do not stop at a division based on the target modification taxonomy, but also consider deviations at different levels of occurrence. For instance, both *mödravårdcentral* and *inflamation* are omissions by, in the first case, a missing 's' and in the latter one an 'm' to obtain the terms *mödravårdscentral* 'antenatal clinic' and *inflammation*. If the aim is to correct omissions, the view of the first deviation as occurring on text level and the second at the substance might be seen as irrelevant. However, in our opinion, it is also valuable to understand the "reason" for the deviation, as in the case of the missing 'm' may result from a so called misencoding where the seeker believes the singleton letter to correctly represent the sound. Considering the omission of the 's', it represents a formal distortion indicating an unclear understanding of the semantics of the intended term. Hence, this deviation occurs at a different "level" than the first one, and this raises the question if it is more important for a portal to be able to manage the first type of deviation, or if it is a first-order deviation which seekers will correct if they are presented with an empty answer list. However, the second case may point to a second-order deviation, thereby an empty answer list causes confusion and possible distrust in the portal's abilities. In other words, not all modifications may have the same consequences considering the trust between seekers and portals, and the need of similar provider attention.

6.5.2 Substance

Substance deviations refer to *graphological* ones, i.e. the way the seeker expresses herself in writing, and contain aspects like *misspellings*, table 6.3, and *mispronunciations*[45] and lead to interpretation problems for the seeker and portal, respectively. However, the division into the different types of misspellings

[45]This type of deviations relates to pronounced and not written communication, hence is not considered in this thesis.

is not completely clear. For instance, dyslexic,[46] confusibles and mispronunciations all refer to similar sounding *graphemes*, i.e. a "feature of written expression that cannot be analyzed into smaller meaningful units" (James 1998: 134).

Deviation		Definition	Example
"Mechanical"	Punctuation	Deviation due to overuse, called *split*, or underuse, called *fusion*, of the "space" between parts of compounds.	hud läkare
	Typographic	Deviation due to *spatial*, e.g. stroke of adjacent key, or *temporal*, i.e. misordered typing of letters.	barnsjukhua sjukgymnats
	Dyslexic	Deviation due to misselection from two letters that can represent the same sound.	brock canser
	Confusibles	Deviation due to misselection of graphemes, often several letters, with similar sound.	förskyld
Proper	Mispronunciation	Deviation due to misselection of grapheme wrongly thought to represent the "target" sound.	epelepsi atros angora fobi
	Written misencodings	Deviation due to misselection of grapheme correctly thought to represent the "target" sound.	colon inflamation

Table 6.3: Types and examples of misspellings.

In the first case, it is when two letters sound the same, in the second when parts of words sound similar, and in the last case when the writing is based on the mispronunciation of a word. In our context, with a given portal vocabulary and possibly a different seeker one, we consider these substance deviations to correspond to ones where a simple look-up of the word would not result in any answers, even though a modification would make the word possible to find, hence interpretable by the portal. In the case of second order mistakes, the portal indicating to the seeker of something possibly misspelt, would not

[46]This term refers to deviations "dyslexics make" (James 1998: 133), but the term may be questioned since the author does not provide any evidence of misselections as using the wrong letter for a sound to be more pronounced among dyslexics. However, we follow James and call this type of deviations dyslexic in its literal sense of a word being difficult to pronounce.

resolve the problem. A rough estimate of the degree of first order misspellings is to study how often a misspelling and its correction are found in the same time interval, where we let a misspelling be a query without answers but within a *Levenshtein distance*, i.e. the minimum number of single character edits, i.e. insertions, deletions or substitutions, required to change one word into an other (Levenshtein 1966), of two from a query in the same interval with answers.

For instance, in the case of our Stockholm County studies, we estimate in the range of 10% of all non-interpretable queries to be due to spelling problems with up to a third being second-order deviations. Hence, an understanding of the types of misspellings which are at higher risk of being second-order is of particular interest, since they may be more prone to impact the seekers' trust in the portals' ability to support the seekers' needs.

6.5.3 Text

Text deviations are results of "ignorance and misapplication of the 'lexico-grammatical' rules of the language" (James 1998: 142) and are divided into *lexical* and *grammatical* problems. The first class consists of formal and semantic errors of lexis, or vocabulary, table 6.4, and the second one of *morphological* (word-level) and *syntactic* ("larger" than word-level) errors of structure.[47] The former group consists of errors due to failure to comply with the rules of word formation, e.g. genitive and tense. For instance, when a noun is formed from the verb *svimma* 'faint', it is written *svimning* and not *svimmning*. Another example is the compound *vårdcentral* 'health centre', misspelt as *vårdscentral*, with the suffix 's' inserted after the first of the two constituents.

Misformations would, according to James (1998: 149), correspond to what will be unknown words to the portal in the case they are considered to be members of the Swedish vocabulary. Consequently, we will by misformations intend the ones not found in either vocabulary, but also the cases when unknown abbreviations of words which are members of the portal vocabulary are used, for instance as in the case of *lpk*, the abbreviated form of *leukocytpartikelkoncentration* 'leucocyte particle concentration'. It is difficult to detect this type of problem, not being able to automatically decide if a word is a member of the Swedish vocabulary, but based on our analysis of non-interpretable queries in chapter 9 we believe the number of misformations to be substantial, since less than 5% of the non-interpretable queries for Västra Götaland County are mappable[48] to UMLS concepts. Hence, this type of problems is to be considered

[47] Since the majority of all queries consist of only one word or arbitrary sequences of words, we will not address syntactic deviations.

[48] We call a query (term) *mappable* with respect to an annotation resource if it has a non-

Deviation		Definition	Example
Formal	Misselection	Deviation due to use of *synforms*, i.e. pairs/triples of words that look and sound similar.	
	Misformation	Deviation due to use of "words" that do not exist in the language.	lpk
	Distortions	Deviation due to "modifications" presented above, e.g. omissions and misorderings.	mödravårdcentral vårdscentral
Semantic	Sense	Deviation due to use of wrong *sense*, e.g. near-synonym, hyponym or hypernym.	husdoktor
	Collocation	Deviation due to misordering words with respect to their "normal" order.	

Table 6.4: Types and examples of formal and semantic deviations of lexis.

in analyses of portal interactions.

Sense deviations are of special interest since they provide insights into how well a portal manages concepts and their neighbourhoods. For instance, in the case of Västra Götaland County, the query *baksmälla* 'hangover' is interpretable, but its hypernym *alkoholförgiftning* 'alcoholic intoxication' will not provide any answers, which might seem irrational to an information seeker. Another type of irrationality in the eyes of an information seeker is if, for instance, the query *aneurysm* is interpretable, but not its synonym *artärbråck*, as is the case for Västra Götaland County.

The occurrence of collocation deviations depends on if the language recognised by the portal is order sensitive, but clearly there should be compounds where the order of the words is of importance. However, we have not been able to identify any in the datasets.

6.5.4 Discourse

Discourse refers to the queries with a focus on the writer's, or seeker's, intention and the reader's, or portal's, interpretation of the queries. The problems at this level can be divided into three types, table 6.5. *Coherence* problems occur

empty annotation in the resource. Moreover, if the mapping takes place without manual support we call it *automatic*.

due to unrelated conceptual relatedness of propositions. From time to time, it may happen that a seeker expresses herself in terms not "appropriate" given the portal's vocabulary, and this type of deviations is called *pragmatic*. Finally, when the "message" of a query is unclear to the portal, it may try to "decode" it by utilising rules like omissions and additions presented above. However, these may in turn lead to, for instance, misunderstandings and this type of deviations is called *receptive*.

Deviation		Definition
Coherence	Topical	Deviation due to proposition not relevant in the context.
	Relational	Deviation due to propositions not conceptually related.
Pragmatic		Deviation due to inappropriate use of linguistic knowledge.
Receptive		Deviation due to problems when trying to decode message.

Table 6.5: Types of discourse deviations.

This, the final level of problems, is the most difficult to study in our setting, since it assumes an ability to understand the intentions of the seeker and portal interpretations. We may assume that the portal's interpretation of the seekers' queries is based on a, more or less, basic look-up of the used query terms in an index mapping to all available answers, i.e. portal pages, where some of these pages are limited to queries associated with certain contextual information like location and used device.

6.6 Queries with many answers

In the previous section, we used the work by James (1998) on second language learning as a basis to study queries without answers, implicitly assuming many of the considered queries to result from information seekers not knowing how to express themselves in the language of "health care". In this section, we focus on the cases where the seeker was able to obtain answers, but possibly too many by expressing herself too vaguely, or the portal facing problems to narrow the seeker's interests. Hence, we could say that the challenge might not only be for the portal to make use of the query and context as such in the best possible way, but also to consider aspects related to the way the seeker expressed herself. For instance, if she uses more general medical terms like *cancer* or specific ones like *cervical cancer*, indicating a more focused interest and possibly also some knowledge of the topic. Another example is if the seeker uses terms, e.g. *vaccination*, which may indicate an interest in preventive actions or ones like *feber hosta* 'fever cough' implicating an interest in treatments. Hence, it is not only important for a health information portal to

consider *what* is typed by a seeker, but also *how* it is done, that is the "style" of the queries.

When studying different types of problems leading to information seekers obtaining no answers, we described deviations at several levels based on the way an information seeker expresses herself (substance), lexico-grammatical patterns (text), and the intentions of the seeker and interpretations of the portal (discourse). Focusing on the information seeker's query, the substance level would then correspond to the actual words used by the seeker, the text level to the grammatical and lexical rules and patterns used to produce the query, and the discourse as described to reflect the intentions of the seeker. With this division as a basis we will describe stylistic features such as the use of *compounds*, *capitalisations*, *acronyms* and *conjunctions* in addition to basics as the number of used query terms to analyse the substance level of queries. Aspects like the type of language and use of lexicological relations like *synonymy* and *hypernymy* of the semantics of the terms is then addressed at the textual level of the queries and answers, and we finish by studying these in the light of the discourse of the interactions.

Each level of stylistic features may tell us something about the seekers and their behaviour, thereby provide insights for portal developers on challenges, and potential solutions. For instance, the use of compounds may lead to portal problems when only the constituents and not the compound are known to the portal. Another example is the use of synonyms where seekers expect the same answer lists, and deviations could possibly lead to distrust and concerns on the "language" to use when searching. The problem in these cases is not that the queries end up without answers, but that a seeker may have to browse through many answer gists before finding one of interest. This is especially challenging in mobile settings where the used devices benefit from being able to present as few answers as possible to the seeker. Hence, we focus on describing some characteristics of queries, called n*-queries*, often leading to more than a given number n of answers for an information seeker to browse before considering an answer interesting, and how these may help information providers decide on answers to present to reduce the effort of the information seekers to find adequate support.

6.6.1 Stylistics

Stylistics is a research area interested in the *expressive means* available and used in written and oral communication. Hence, studies on (linguistic) deviations as studied in the previous section may be seen as an instance of stylistics. Crystal (1970: 99) labelled stylistics a topic that "covers the whole complex

of varieties and styles that make up 'a' language – comprehending such differences as the distinction between [...] formal and informal, scientific and religious, and many more". Or paraphrasing Crystal (1970: 100) "an author can bring into his writing any use of language he pleases", but are there common linguistic features to be found among authors made up of health information seekers, and how can a portal use them to improve its ability to support the needs of the seekers? Furthermore, how might this variety of language relate the used expressions, i.e. queries and answers, and the situations where they are used? We will in this part study the existence, and features of *stylemes*, that is, linguistic units marking a certain expression of the users. For instance, when an information seeker types a medical term it may convey information of an interest in, and possibly knowledge of, medicine, which cannot be said of the use of plural suffixes. Hence, stylemes "tell" something about the seeker, more than her ability to use certain query terms and their properties.

6.6.2 Substance

When we studied deviations in the previous section, they were made up of those related to misspellings resulting from "mechanical" problems, such as punctuation and typographic ones, and proper misspellings possibly originating in lack of proper knowledge of the used language, e.g. Swedish spelling of *colon* as *kolon* or the use of grapheme "ng" wrongly thought to represent the "target" sound "g" in *angora fobi* (intended expression was *agorafobi* 'agoraphobia').

When considering queries with obtained answers, the substance level will not deal with misspellings, but with *spelling* patterns and variations as the use of capital initial letters to indicate places and compounds to refer to health care units and professions. Hence, we will study the existence and use of substance stylemes of health information interactions as evolving in the use of health portals like 1177.se and vardguiden.se. Of particular interest are the ones where stylistic differences may be seen between n-queries and queries in general possibly satisfying the seeker's needs.

One of the most obvious aspects to study, used in many different research areas on information search, is the *query length* and if there is a difference between the domain of interest and information search in general. If so, this may indicate a different language or use thereof. For instance, in English medical terminology diseases like *bröstcancer* would be expressed as *breast cancer*, hence with a space between the constituents of the compound, and this difference would be more marked in domains where compound expressions are more common than in general. From a portal perspective, the use of compounds is

interesting since it raises the question on how its vocabulary should be defined, e.g. if *bröst cancer* should be treated as a synonym of *bröstcancer*. Another example is the increased use of interactive portal suggestions of queries in domains where the language may be more challenging or the portal wants to better *narrow in on* the seeker's interest to provide better support. In the context of n-queries, the question of interest becomes if and how the query length relates to the ability to satisfy the information seeker's needs.

As discussed above, the use of compounds may differ between languages, but the ability to "construct" them impacts the interaction between information seekers and portals. That is, if the used language and domain of interest, e.g. Swedish health information search, is more prone to compounds the portals will have to consider this in the interpretation of queries. For instance, in our deviation studies in chapter 9 we note compounds related to health care units constructed from a profession or activity and terms reflecting a place, e.g. *vårdcentral* 'health centre' being a compound of *vård* 'care' and *central* 'centre' or *barnmorskemottagning* 'antenatal clinic' of a form of *barnmorska* 'midwife' and *mottagning* 'clinic'. One could hypothesise that compound problems would be independent of n for n-queries in cases where the query was indexed by the portal, since compounds per definition are used to constrain a head stem of the compound. For instance, *mottagning* 'surgery' results in more, and possibly different, answers than *barnmorskemottagning* 'antenatal clinic'. Hence, if the latter is a member of a set of n-queries it is indexed by the portal, only further terms would constrain the area of interest.

In addition to searching for professions and health care units, it is common in the setting of health information search, as discussed in chapter 9 on queries without answers, to constrain the search to certain locations. For instance, to be interested in emergency units as close as possible to home. Since locations are often written with a capital letter, an interesting question is how *capitalisation* is used and if this may be used by an information portal to provide better seeker support, especially in the case of n-queries.

A fourth substance feature of potential interest is the use of *abbreviations* and *acronyms* since in medicine, especially considering laboratory and diagnostic procedures, these are commonly used, for instance, as seen by the queries *lpk* (leukocyte particle concentration) and *kbt* (cognitive behavioural therapy) in chapter 9 (tables 9.8 and 9.15). Moreover, by these often being short fragments of letters, a portal index based on parts of words may unintentionally index these as if they were acronyms or abbreviations.

A final aspect which may be of potential interest in the setting of information search is the use of words like *och* 'and' as search *conjunctions* indicating an interest in both its constituents and expressions like the multiword expression *sex och samlevnad* 'sex and relations' referring to a Swedish phrase used

to denote topics related to sexuality and rights. Yet again, the underlying index may treat this phrase as a sequence of three terms or as a single entity, possibly with more pronounced impact in the case of n-queries.

To summarise, an analysis at a substance level of queries deals with stylistic aspects related to the use of, for instance, query length, compounds, capitalisation, abbreviations and acronyms and the use of conjunctions to better understand health information search as a form of human–computer interaction and to utilise this knowledge to improve the support to information seekers.

It is not only queries which may be of interest, and a similar analysis as the one above can be done to characterise the answers, especially those related to *n*-queries. Hence, we may study the substance of answers corresponding to the actual portal pages and their organisation in hierarchical structures. For instance, at the top level of the portal we may find paths to pages related to facts on diseases and treatments, health care management and links to different care givers. At the next level we may find a division of the facts pages into ones related to different types of diseases, body parts or into thematic areas like pregnancy and old age. Hence, the click itself, reflecting this organisation of pages, becomes an entity which may be treated in a similar way as queries. For instance, a click like `Egenvardsguide/Sjukvardsradgivningen/?CatId=28356ChapId=28357` indicating an interest in a specific chapter and category of *Sjukvårdsrådgivningen* 'health care advice' and *Egenvårdsguide* 'self-care guide'. By this, we can say that an answer, or click, consists of several levels of associated information and each part of the answer points to this information. Consequently, we will by *answer level* n of an answer refer to the pointer and associated information at the n-th level from the top. For instance, answer level two in our example corresponds to *Sjukvårdsrådgivningen* 'health care advice' and its related information, given its context *Egenvårdsguide* 'self-care guide'. Hence, the *answer type* is health care advice with facts on self-care guidance.

6.6.3 Text

In the previous section we looked at "surface" aspects of the interaction between information seekers and portals with a focus on the seekers' way of expressing themselves. However, the reason to type certain queries is for a seeker to try to convey an interest, or need, to the portal and expecting it to react in a feasible way. Hence, a substance analysis of an interaction may benefit from a better understanding of the *semantics* of the used terms. For instance, if they belong to common language or are domain-specific expressions, and if there is a preference to use specific or more general terms. The first question is related

to the vocabularies of the seekers and portals (section 6.5.1), and the second one to the "level" of detail a seeker may expect in the answers provided by a portal.

In the case of medical terms related to, for instance, diseases, body parts and treatments there is a tradition, also implemented in the UMLS and many of its sources, to organise the terms according to different criteria. For instance, they may be related in an *onomasiological* (meaning-to-form) manner according to if they are *hypernyms* or *hyponyms*. Hypernymy describes that something is a "type of" something else, like *cervical cancer* is a type of *cancer*, and hyponymy reflects the "inverse" of hypernymy. Hence, these types of relations describe a lexical cohesion and semantic coherence among query terms. The most common relation of this type which a portal should manage is *synonymy*, that is, terms with the same, or very similar, meaning. For instance, *nackstelhet* 'neck stiffness' and *stel i nacken* 'stiff in the neck' both refer to the same medical concept *Neck Stiffness*. The other perspective of terms are to view them in a form-to-meaning, called *semasiological*, perspective focusing on the aspects like *polysemy*. For instance, *sjukhus* 'hospital' can denote the building as such or a health care unit. A common theme among these relations is that they invite to interpretation problems, but also opportunities, for a portal when used by information seekers.

To summarise, as in the case of stylistic aspects on a substance level, there are many different ways to view queries at a text level, but a common theme is the use of their semantic interpretations and to study relations among the terms based on these.

Considering answers, they may be treated in a similar way as queries, but in this case each (sequence of) level is assigned a concept and a semantic type. Thereby, each answer is turned into a sequence of concepts. For example, the answer `Egenvardsguide/Sjukvardsradgivningen/?CatId=28356ChapId=28357` results in the semantic sequence (*Inquiry*, *Disease*, *Self-care*, *Advice*). This annotation can be used in analysis of the interests of information seekers and is especially useful in a discourse analysis of relations between queries and answers.

6.6.4 Discourse

The discourse level addresses in our context relations among the queries, answers and clicks considering relations both at the substance and text level. For instance, addressing questions like if an information seeker posts the query *feber hosta* 'fever cough', is she then more interested in answers related to treatments or to preventions. But it is also interesting to study how a por-

tal treats queries with well-defined names, such as pharmaceuticals and body parts.

6.7 Properties of a trustworthy portal I

A common theme in this thesis has been the importance of a health information portal to establish and maintain seekers' trust. In section 4.2 we saw how there are several different types of trust, all leading up to the notion of a trustworthy behaviour. Hence, if a portal acts in a trustworthy way in the eyes of an information seeker, she is willing to "depend" on the portal's advice and information.

An obvious, and non-trivial, question is how a portal may live up to this, only having access to the seekers' current and past queries, contexts and the portal's vocabulary and preference relation.

First of all, we addressed the topic of the importance of the context as both a constraining and focusing aspect of a query. We concluded in the discussion on queries without answers that a substantial number of incomplete search rounds seem to originate in a possibly too emphasised use of context information, especially considering locations. In section 9.2.2.1 we will elaborate on the use of preference relations which are rational and takes into account both location- and time-dependency based on the simple principle that whenever the interest is to obtain answers referencing health care units the preference should be to present ones that are as close as possible to the seeker, but also open at the time of search. Hence we can state our first principle for a trustworthy portal as follows:

Principle 1: If a seeker's query indicates an interest in health care units, the portal's rational preference relation should take into account location and time information to promote answers referencing units with high availability given the seeker's location and time of search.

The next addressed area of interest was the portal's treatment of linguistic deviations among seekers' queries leading to incomplete search rounds, and using the framework of error analysis we are able to divide the types of answers into different groups and levels. Chapter 9 is devoted to this type of challenges, and we can state the following principle:

Principle 2: If a seeker's query is a non-interpretable possible second-order deviation, a portal's rational preference should try to categorise it in accordance with James' framework and apply adequate measures to "correct" the deviation and make the query interpretable.

The second principle targeted the problems with queries ending up without answers, and the third topic of interest, the focus of chapter 10, is the cases when the seekers end up having to browse several answers before finding one of interest. In this case the tool is the use of a stylistics analysis to ensure rational preference relations trying to minimise the number of needed answers, and we propose the following principles:

Principle 3: A rational preference relation should respect onomasiological properties like the same answer lists for synonyms and a sensible treatment of hypernym and hyponym relations among queries.

Principle 4: A rational preference relation should treat semasiological properties like polysemy in a sensible way to avoid the "wrong" answer to the "right" query.

Further principles and instances of the ones above will be introduced at the end of the studies of queries without answers (section 9.5) and queries with many answers (section 10.5), and then summarised and discussed in chapter 11.

Part II

Case study

7 INTRODUCTION

We began this work by a hypothetical example where an information seeker with a stiff neck and fever who, using a Swedish health portal, wished to find out more on potential causes of her problem and whether to seek care. We also elaborated on how the portal may have acted on the provided query and available context information, and in this part of the thesis we will present how a theoretical model along the lines introduced in the first part of the work can be used to analyse seeker and portal behaviours.

In our analysis we use three datasets comprising usage of the Swedish health portals 1177.se and vardguiden.se for four counties in Sweden (Stockholm,Västra Götaland, Östergötland, Jämtland), four parts of Stockholm city (Hägersten-Liljeholmen, Rinkeby-Kista, Skarpnäck, Östermalm) and one for the county of Västra Götaland. Each dataset is chosen based on the type of registered information and demographic aspects like age, education, proportion of immigrants and measures of health problems in the considered parts of Sweden.

With the help of the provided data and theoretical framework, many avenues of analysis with a natural language perspective could be chosen, none to our knowledge previously followed, to better understand how official Swedish health portals are used and what to be learnt for future portal solutions. In this work we take the first steps towards this, and focus on two areas of importance to information seekers – when *no answers* to queries are obtained and when *many answers* are provided. In the latter case, we are especially interested in the cases where seekers in general have to browse more than five answer gists before finding one of interest. Hence it is not the number of answers per se that is of interest, but how difficult it is for a seeker to find the "right" one. These areas are of interest from a portal provider perspective, since ultimately a portal should be able to provide few, but relevant, answers to the search queries, based on its interpretation of the queries and their contexts.

As will be presented in chapter 8 on material and methods, a substantial number, possibly even in the range of 5%, of queries result in empty answer

lists – incomplete plays in our search game terminology. In our opinion, based on the role of public health portals in society (chapter 2), this is a problem potentially jeopardising the aim of these portals, but also a challenge to be addressed by the portal providers. This is the topic of chapter 9, where we try to characterise queries without answers.

When seekers obtain answers, there are often many different types of answers, e.g. related to treatments or preventions, but often a seeker's interest is limited to a specific one of the covered areas of answers. The challenges of one given query having several answers are addressed in chapter 10, where we also show how semantic information may help a portal constrain its answers based on the semantic information in a query and its context, that is, how possible intentions induced by seeker moves may provide insights into feasible portal actions.

In chapter 11, these different aspects of information search are tied together in a number of principles we believe a trustworthy (health) information portal should possess.

As elaborated on, with a basis in our theoretical framework, the main challenge for any portal in the service of the public is to establish and maintain trust in its ability to satisfy the needs of information seekers, where the means to achieve this is limited to interpretation of written utterances and adequate answers. Hence, the trust needs to have a basis in the dialogue, or interaction, among the seeker and portal. For instance, a portal which is not able to provide any feedback when a seeker is interested in smallpox, or when more than 10% of the people interested in borrelia are not sure how to spell the name of the disease, will be at risk of distrust. Similarly, if the portal provides answers on prevention of a disease when the seeker, in her own opinion, clearly shows an interest in treatment of the disease, or if the information, according to the seeker, is irrelevant and concerns other problems. Finally, it is important to stress that this area of research is vast and we are only able to provide a few reflections on the topics in the setting of official Swedish health portals, and to introduce a theoretical framework to allow future discussions on preferences, rationality and trust among two equally important and active participants in the sense of the information seeker and portal.

8 MATERIAL AND METHODS

In this chapter we present the *material* (section 8.1), i.e. the search logs and annotation resources, making up the basis of our case study, and the *methods* (section 8.2) used to induce and analyse search games based on these. In our presentation we rely on the terminology and discussions introduced in part I on our search game framework. The methods are divided into ones related to *normalisation* of the data like management of cases to allow for a consistent analysis, *annotations* to add additional information like semantics, or meaning, to the queries, and methods used to *analyse* the data.

The search logs contain different interaction and context data captured by the public health portals 1177.se and vardguiden.se during the period June 2010 – September 2013. As presented in section 8.1.1, the type of context information captured has changed over time, but also at county and portal level changes have been implemented. Hence, some (subsets) of the data are more useful than others for certain research questions. In addition to the search logs, information on medical concepts and their relations (section 8.1.2) is used to organise and analyse the log data.

8.1 Material

The material used in this case study is divided into *search logs* and *annotation resources*, and each of them is described in detail in this section. As discussed in chapter 6, the contents of the search logs are divided into the *queries*, their *contexts* and data related to the *answers*, and in section 8.1.1 we present an overview of these for the considered search logs. This part also addresses the challenges of change over time in which information is captured, and the consequences for our study. We also elaborate on the problems of inconsistent session information for the interactions.

In section 8.1.2, we present the resources used to annotate the search log entries, i.e. the UMLS sources and the Swedish Snomed CT database. When resources are used to add meaning to terms, we call them *semantic resources*.

8.1.1 Search logs

Table 8.1 presents a general overview of the search logs, and their contents and use in this thesis. For basic statistics on the search logs, see tables A.1–A.5. The data used in the case study was provided by Findwise and Euroling AB,[49] in agreement with Stockholm County Council (J Bjurel), as anonymised records without any ability to trace the seekers in any other way that in some cases the interactions in a session were grouped and contextual information like time and location provided. Hence, reasonable steps have, in our opinion, been taken to ensure seeker privacy, and by the types of analyses presented no group of individuals at a more fine-grained level than parts of a city or county is distinguishable.

Dataset	**Stockholm**	**Sweden**	**Västra Götaland**
Source	Vårdguiden	Vårdguiden	1177
Time	Oct 2010 – Sep 2012	Sep 2012 – Sep 2013	Jun 2010 – Sep 2011
Location	Stockholm regions	Swedish counties	Västra Götaland
Device	Mobile Non-mobile	Unknown	Unknown
Domain	Article Blog FAQ Health Unit	All Health Unit	Unknown
Answers	All domains	Chosen domain	Unknown
Click	No	No	Yes
Rank	No	No	Yes
Interval	5-min	5-min	Session

Table 8.1: Overview of the available datasets.

The data can be described along the context dimensions of geographical *location*, used *device* and *domain* of interest for answers in addition to *time*-related information. For instance, Vårdguiden logs for the interval October 2010 to September 2012 are useful when considering questions related to used device, number of answers per domain and different demographic aspects traceable at the level of different parts of the Stockholm County.[50] In the following sections, we discuss in detail these different aspects and their implications for our analyses. In the rest of this work, when referring to any

[49] findwise.com, siteseeker.se

[50] This dataset has also been presented in (Kokkinakis and Eklund 2013).

of the three *datasets*, we do so by the names Stockholm, Sweden and Västra Götaland. Moreover, each dataset is further divided into segments reflecting different context constraints, e.g. location or used device, see below.

As of November 2013, the two portals have been merged into one called 1177 Vårdguiden, sharing interface and search engine. In this work we treat the portals as different entities, since they in the past had partly different interfaces and search engines, reflected by the search logs. However, in the studies we may refer to the new merged portal, but this will then be emphasised.

One of the most challenging aspects of the raw material has been the different ways the datasets treat the concept of session, thereby which information is logged and how. A session is intended to be a well-defined set of interactions initiated and ended by a given information seeker. Hence, it may contain only one query without any answers up to many queries and clicks over a longer time interval. However, it is important to note that the notions of sessions, utterances and search rounds are challenges to deal with in all analyses, since the logs do not contain reliable identifiers to differentiate between seeker and technically induced log entries in the majority of the logged interactions. Table 8.2 presents how we deal with these notions for the different datasets, and each of them is discussed in detail in the rest of this section. It is important to note that for Stockholm the number of distinct utterances per interval is equal to the number of search rounds, which may not be the case for Sweden and Västra Götaland.

Dataset	**Stockholm**	**Sweden**	**Västra Götaland**
Interval	All entries within an interval of 5 min	All entries within an interval of 5 min	Sessions given by log
Context	Location Device	Location Domain	No
Utterance	Only considered once within interval	Each occurrence	Each occurrence
Answer	Max answer list size per utterance within interval	Each occurrence of answer list size	–
Click	–	–	Each occurrence
Search round	Utterance – answer pair	Utterance – answer pair	Utterance – click pair

Table 8.2: Approaches used to address challenging dataset notions.

To clarify these aspects in the setting of the Stockholm dataset, table 8.3

presents an example where the query *urinvägsinfektion* 'urinary tract infection' was posted to Vårdguiden in October 2010. In this case the query was logged together with the number of answers for three search domains. However, we may assume that the query was only posted once, but we do not know which domain that was of interest to the seeker. For this type of log entries the query is only counted once, and the number of answers is assumed to be the maximum of the sizes of the lists of answers. Hence, in this case we have one query resulting in 41 answers, but without any information on domain of interest or if the seeker clicked any answer. Moreover, these three entries are only counted as one utterance with the query *urinvägsinfektion* and context comprising any potential location and used device. Since we do not have any information on domain of interest these three entries will also be viewed as one complete search round with 41 answers.

Session	Time	Query	Answers	Domain
1	2010-10-01 04:58	urinvägsinfektion	41	Article
1	2010-10-01 04:58	urinvägsinfektion	7	Health Unit
1	2010-10-01 04:58	urinvägsinfektion	1	FAQ

Table 8.3: Example of search log entries for Stockholm.

Table 8.4 shows a similar example posted in 2013 from Uppsala, but in this case we know that the seeker actively posted the query twice by first looking at the outcome of a general query followed by looking at the results when the domain of interest was limited to health care providers in the Uppsala region. In this case the seeker is considered to have posted one query with different domains of interest, hence two utterances of the query with a context comprising domain of interest and location. Since each utterance resulted in non-empty lists of answers, we have three complete search rounds.

Session	Time	Query	Answers	Domain	Location
2	2013-01-16 10:00	urinvägsinfektion	5	All	Uppsala
2	2013-01-16 10:45	urinvägsinfektion	1	Health Unit	Uppsala

Table 8.4: Example of search log entries for Sweden.

Västra Götaland is the dataset with, in our opinion, the most accurate session information and table 8.5 presents the estimated distribution of *session lengths*, i.e. the number of complete search rounds (query and click) carried out by a seeker during a session. As almost 50% of all sessions consist of only one round, our analyses and discussions will focus on this type of rounds corresponding to ones in a sequential static search game. However, in the case

of Västra Götaland the problem is that we do not know anything about search rounds not resulting in any answers or where the seeker decided not to click any answer among the provided ones.

Interval length	Intervals
1	13,172
2	5,877
3	3,249
4	2,203
5	1,064
>5	2,316

Table 8.5: Distribution of interval lengths for Västra Götaland.

To summarise, due to different type of information captured by the search logs, each dataset has a specific way of counting utterances and search rounds.

8.1.1.1 Interval

For Stockholm and Sweden no reliable session information is provided. Hence, to study topics related to estimates on the occurrence of different types of queries, we divide the log into non-overlapping 5-minute *intervals* as a rough substitute for sessions. The choice and use of intervals is further discussed in section 8.2.4.2.

The difference between sessions and intervals is clearly seen if we compare the session length from Västra Götaland, table 8.5, and *interval length*, i.e. number of complete search rounds, for Sweden, figure 26. For Västra Götaland almost half of the sessions where an information seeker chose an answer contain only one query and one answer. Considering the Sweden interval lengths, it shows a completely different pattern, thereby disqualifying interval-based datasets from detailed seeker–portal interaction analysis.

8.1.1.2 Query

By figures 27 and 28, the search log for Stockholm shows a trend of fewer search rounds per month, and a peculiar pattern with a major reduction in the number of rounds per month during the end of 2010 and beginning of 2011 not found for Västra Götaland, figure 29. However, also for Västra Götaland we see a slight trend towards fewer rounds per month from early 2011.

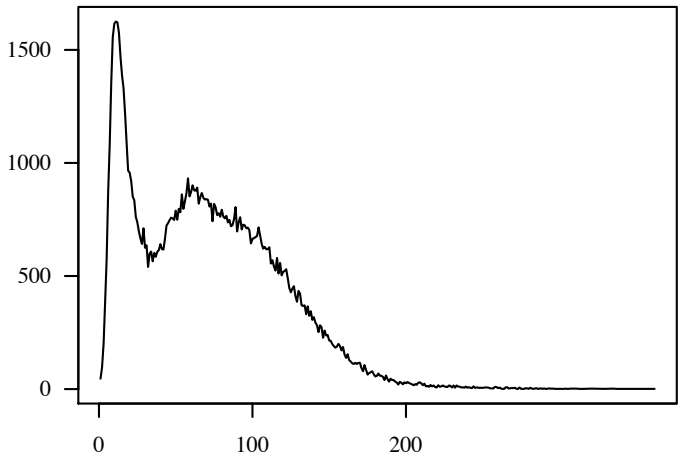


Figure 26: Distribution of interval lengths for Sweden.

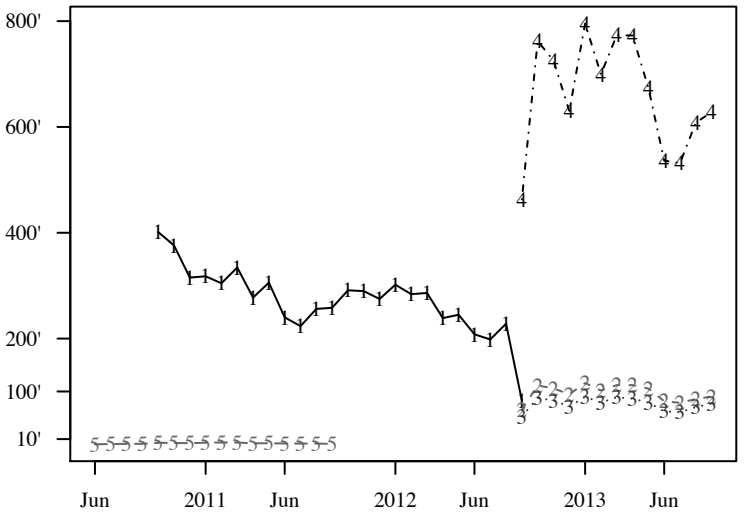


Figure 27: Number of search rounds per month for Stockholm (1), Stockholm County (2), Västra Götaland County (3), Sweden (4) and Västra Götaland (5).

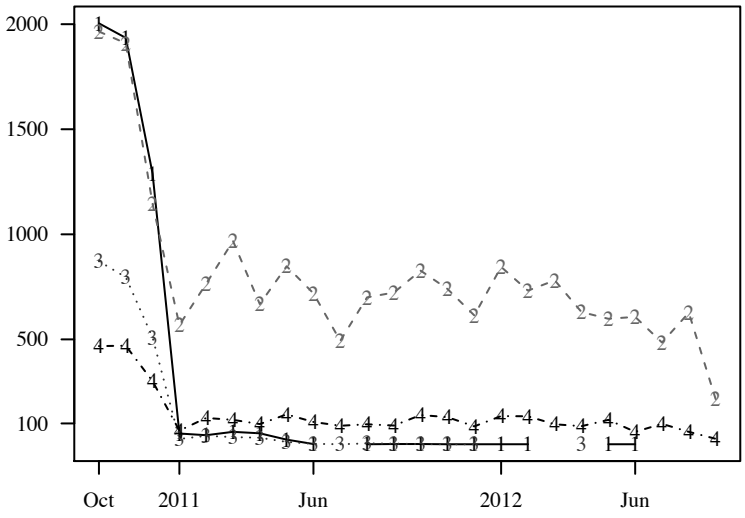


Figure 28: Number of search rounds per month for selected parts of Stockholm; Hägersten-Liljeholmen (1), Östermalm (2), Rinkeby-Kista (3), Skarpnäck (4).

Analyses of other parts of Stockholm show a similar pattern, which cannot be explained. Thereby, we decided to constrain our in-depth analysis of Stockholm to the interval October–November 2010. Unexplained trends, as in the case of Stockholm, are not found for Sweden, figures 27 and 30. Hence, for analyses utilising this dataset no time interval limitations are made. Worth noticing is a tendency of fewer search rounds per month for all counties during the second half of 2013 in comparison to the first half, and similar patterns of number of rounds per month during the span of the log. Moreover, since the search logs are similar also for September 2012, we do not exclude this month from the analyses, even though there was a change in logged information during this month.

Another interesting aspect reflected in the claim by Stockholm County Council (SLL 2013) that Vårdguiden Stockholm had two million visitors in January 2013 in comparison to our estimate of 89,932 search rounds, is that definition of concepts like visits, queries and sessions is a challenge, with potential consequences for proceeding analyses.

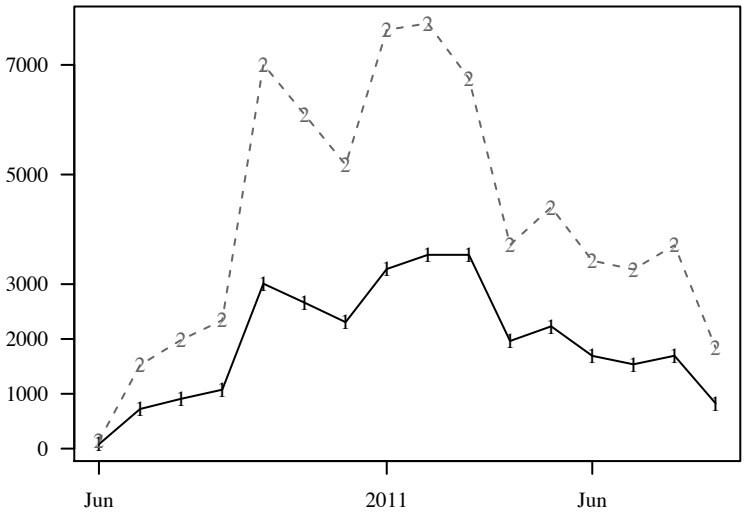


Figure 29: Number of search rounds per month (1) and complete rounds per month (2) for Västra Götaland.

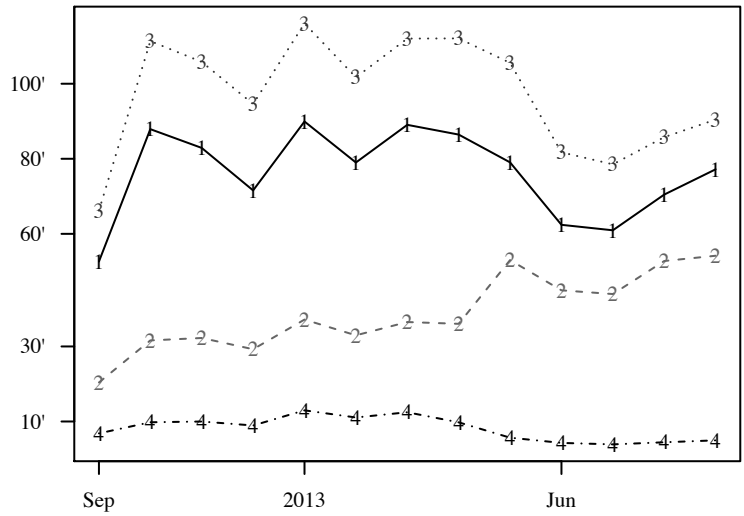


Figure 30: Number of search rounds per month for Stockholm (1), Östergötland (2), Västra Götaland (3), Jämtland (4) Counties.

8.1.1.3 Location

The way the searches are constrained by location differ among the datasets. For Stockholm the most coarse-grained location is Stockholm County itself, but the log also contain constraints like City or Southern Suburbs with unknown limits and use. The most fine-grained division of Stockholm, and the one of main interest to us, is into the boroughs of the county, for instance Östermalm and Rinkeby-Kista. As seen by figure 28, for the presented boroughs the number of search rounds per month show similar trends, especially for the period October–November 2010, figure 31.

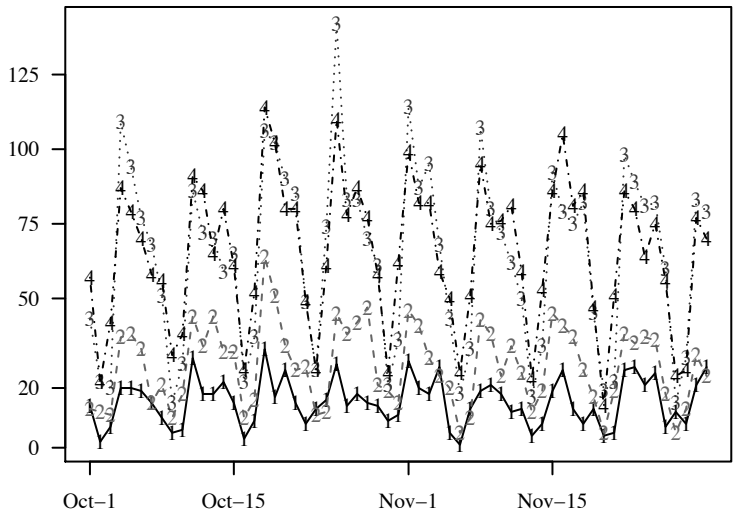


Figure 31: Number of search rounds per day during October–November 2010 for selected parts of Stockholm; Skarpnäck (1), Rinkeby-Kista (2), Östermalm (3), Hägersten-Liljeholmen (4).

Considering Sweden, "any Swedish county" is the most general location filter, herein denoted Sweden (unspecified county), but the normal constraint is at county level. This is set by default by the portal utilising the IP-address of the seeker, or "cookies" from previous searches. However, when a search is constrained to Health Unit it is also possible for the seeker to further constrain the location to cities in the county. As seen by figure 30, the number of rounds per month show, in general, similar trends for the presented counties.

For the non-mobile part of Stockholm we can also differentiate between locations in the Stockholm County used to constrain the information searches and table 8.6 presents a summary of the interactions for our selected locations

during October–November 2010. As discussed below, we decided to limit our interest to four parts of Stockholm making up 9.0% of all Stockholm search rounds spread over 43.5% of all intervals for the considered period of time. To note possible differences caused by demographic aspects, the boroughs Rinkeby-Kista, Hägersten-Liljeholmen and Östermalm (table 8.6) were chosen based on aspects like *age* of the population, highest level of *education*, *health status* and *country of birth*, see discussion in section 8.2.1 on method used to choose samples. In addition to these, Skarpnäck was chosen as a representative of an area where more than 50% of the search rounds seem to not result in any answers (table 9.2), hence being incomplete.

From the Sweden dataset the counties Stockholm, Östergötland and Jämtland, were selected based on similar criteria as for Stockholm, table 8.7 and figure 30, see section 8.2.1 for details.

Location	Population	Intervals (%)	Search rounds (%)
Skarpnäck	44,608	5.9	0.8
Rinkeby-Kista	47,872	10.1	1.4
Östermalm	67,147	21.9	3.3
Hägersten-Liljeholmen	78,826	20.6	3.4
Total	238,453 (11.2%)	6,351 (43.5%)	10,418 (9.0%)

Table 8.6: Distribution of non-mobile searches for selected locations in Stockholm October–November 2010.

Location	Population	Intervals (%)	Search rounds (%)
Jämtland	126,201	38.3	1.4
Östergötland	433,784	79.7	6.8
Västra Götaland	1,600,447	92.6	16.9
Stockholm	2,127,006	90.9	13.3
Total	4,287,438 (44.9%)	107,286 (98.2%)	2,861,434 (38.4%)

Table 8.7: Distribution of searches in selected counties for Sweden (percentages only based on utterances and intervals with county information).

8.1.1.4 Device

Since we are able to consider used device[51] for Stockholm, but without any further location information, we have that 83.9% of all intervals contain mobile

[51] A *mobile device* refers to a used search device considered by the portal to be mobile, based on internet protocol address, seeker or portal interface.

interactions and that the mobile search rounds make up 10.0% of all rounds. Figure 32 presents these over time, but as discussed in section 8.1.1.2, the Stockholm dataset shows some patterns over time which we are not able to explain, and we do not want to claim the presented increase in mobile rounds over time as a confirmation of a trend. Still, we will use the complete dataset whenever discussing aspects related to mobile searches.

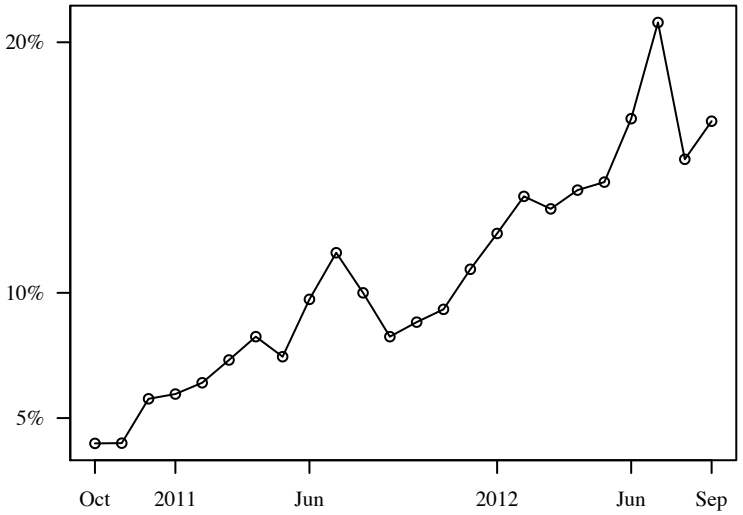


Figure 32: Proportion mobile search rounds per month for Stockholm.

8.1.1.5 Domain

By the configuration of the portals, the searches are divided into different domains depending on the type of portal page from which the queries are posted. For instance, when a query is posted from the main page, the domain is generally called 'All' in the case of the Sweden dataset and 'Article' for Stockholm, and no information is used to constrain the answers.[52] However, if the query is posted from a so called 'Health Unit' page, the portal assumes the seeker to be interested only in information on health professions and care givers, for instance location and opening hours. In these cases, the search is possibly further constrained to a location. The other domains of interest available for Stockholm searches are 'Blogs' and 'FAQ'. In Sweden, the searches are in

[52] In the case of Stockholm, the number of answers for the other domains may, for unknown reasons, not sum up to the ones for the unconstrained Article answers.

some cases further constrained to cities in a given county when the domain of interest is 'Health Unit'.

The logs have captured data in slightly different ways with respect to this aspect, by Stockholm registering number of answers for each domain independent of seeker choice, but in the case of Sweden only information for considered domain is logged. For Västra Götaland no information on domain is captured. Consequently, only Sweden can be used when studies related to domain of interest are considered, and Stockholm when answers per domain is the focus.

8.1.1.6 Answers and clicks

As discussed above, the Stockholm dataset contains the number of answers for all available domains, but for Sweden only the answers for the chosen domain are presented. However, in neither of these cases any further information than the number of answers is provided, and only Västra Götaland contains information on the answers chosen by the seeker, the clicks, and the ranks of the clicked answers. In this case no information on queries without clicks is available. Moreover, the rank does only indicate the position in a list of answers corresponding to a page of answers. Hence, if we assume 20 answers per page the rank 10 could be an indication of answer 10, as well as 30, 50 etc.

The click information for Västra Götaland is logged as an internet address and as discussed in section 8.2.4.3, this address is manually categorised using UMLS information as exemplified in table 8.8. The result is one or more semantic types describing the click, with associated concepts as the representative of the type. Thereby, we obtain a description of the clicks which can be related to other data, e.g. queries, mapped to the UMLS.

8.1.1.7 Choice of samples

Based on the discussion above, table 8.9 summarises the chosen samples used in this work. The samples will be referred to as Stockholm Non-mobile, Stockholm Mobile, Sweden and Västra Götaland in analyses and discussions. By the Stockholm dataset we intend the union of the mobile and non-mobile samples.

8.1.2 Annotation resources

Given the queries and context information captured in the logs, we will base many of our analyses on the ability to add information, or *annotate*, describing

	Level 1	Level 2
Click	sv	Fakta-och-rad
Concept	Question (inquiry)	Advice
Semantic type	Intellectual Product	Health Care Activity
Answer type	Inquiry	Advice
	Level 3	**Level 4**
Click	Rad-om-lakemedel	Lakemdel-och-muntorrhet
Concept	Pharmaceutical Preparations	Xerostomia
Semantic type	Pharmacologic Substance	Sign or Symptom
Answer type	Drug	Sign or Symptom

Table 8.8: Example of mapping click to UMLS concept and semantic type for Västra Götaland.

different aspects of the data as provided by other information resources. We are especially interested in the use of the Unified Medical Language System, abbreviated UMLS, by the US National Library of Medicine (2014), to obtain categorisation and relations of concepts expressed as queries and answers in our logs, for instance, to add information on diseases, symptoms, treatments and administrative terms, thereby allowing analysis at a different level than a purely morphological one. In other words, the sources allow us to add *semantic* information and relations as synonymy to the queries.

The semantic resources mainly used in this thesis are the Swedish Systematized Nomenclature of Medicine Clinical Terms, abbreviated Snomed CT,[53] (Socialstyrelsen 2013) and Medical Subject Headings, MeSH, (NLM 2013b), and relations derivable between these found in the UMLS.

8.1.2.1 The Unified Medical Language System

The Unified Medical Language System, produced by the US National Library of Medicine, is a set of resources aimed at integration and distribution of "terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records" (NLM 2014).

The UMLS contains three tools, called "Knowledge Sources". These are the Metathesaurus containing terms, concepts, etc, the Semantic Network containing semantic types and relations, and the SPECIALIST Lexicon and Lexical

[53] Swedish Snomed CT is not part of the officially distributed UMLS dataset, but has to be managed separately, see section 8.1.2.2 on Snomed CT.

Dataset	**Stockholm Non-mobile**	**Stockholm Mobile**	**Sweden**	**Västra Götaland**
Source	Vårdguiden	Vårdguiden	Vårdguiden	1177
Time	Oct 2010 – Nov 2010	Oct 2010 – Sep 2012	Sep 2012 – Sep 2013	Jun 2010 – Sep 2011
Location	Skarpnäck Rinkeby-Kista Östermalm Hägersten-Liljeholmen	Unknown	Jämtland Östergötland Västra Götaland Stockholm	Västra Götaland
Device	Non-mobile	Mobile	Unknown	Unknown
Domain	Article Blog FAQ Health Unit	Article Blog FAQ Health Unit	All Health Unit	Unknown
Answers	All domains	All domains	Chosen domain	Unknown
Click	No	No	No	Yes
Rank	No	No	No	Yes
Interval	5-min	5-min	5-min	Session

Table 8.9: The chosen samples of the datasets.

Tools, which are natural language processing tools. In this work only the first two are used.

In the Metathesaurus, every *term* from the terminologies is connected to a *concept*, which links different *names* for the same concept from different vocabularies. For example, the terms *Diabetes Mellitus, Type 2*, *Diabetes Mellitus, Type II*, *type 2 diabetes* and *NIDDM* all refer to the same disease and are connected to the same concept. Every concept has a unique *concept identifier*, which links all concept names, relationships and attributes related to a particular concept. In the Metathesaurus there are more than two million concepts, and it contains more than 100 source vocabularies, some of them translated into different languages.

Every UMLS concept is assigned a *semantic type* organised in the Semantic Network in a hierarchy containing different levels of specificity, where the semantic category chosen for a concept is as specific as possible. For instance, diseases (e.g. influenza and diabetes) are mostly mapped to the semantic type *Disease or Syndrome*, but for some diseases there is a more specific subcategory, e.g. cancer diseases which are mapped to *Neoplastic Process*. The Semantic Network contains 133 semantic types and relations between the types. The hierarchy of semantic relations contains 54 different relations, such as *isa*, *treats*, *causes* and *contains*.

In this work we use UMLS version 2013AB (released November 2013) and the available terminologies of restriction categories 0–3 and 9.

8.1.2.2 Systematized Nomenclature of Medicine Clinical Terms

One of the vocabularies in the UMLS is Snomed CT (Systematized Nomenclature of Medicine Clinical Terms), which is a clinical terminology used in, for instance, electronic health records to link clinical information. It is owned by the International Health Terminology Standards Development Organisation (IHTSDO), and has been translated into many languages, including Swedish.

Snomed CT contains over 300,000 concepts, and more than 285,000 of these have been translated into Swedish. The concepts are organised in 19 hierarchies, for instance *Clinical finding*, *Body structure* and *Substance*.

The National Board of Health and Welfare in Sweden is responsible for the Swedish translation of Snomed CT as part of *National Strategy for eHealth* (Socialdepartementet 2010) to facilitate communication in health and social care, including statistics, research and education. The Swedish Snomed CT is not included in the UMLS. Licences are provided by The National Board of Health and Welfare in Sweden (Socialstyrelsen 2013). The version used in this work is the Snomed CT version 2013-01-31, and the Swedish release ver-

sion SCTSE_2013_2. The Swedish Snomed CT is aligned with the US version found in the UMLS by sharing the same Snomed CT identifier for semantically equal concepts. To simplify the presentation we will by UMLS refer to the UMLS with the addition of the Swedish Snomed CT.

8.1.2.3 Medical Subject headings

MeSH (Medical Subject Headings)[54] is another one of the terminologies included in the UMLS, used for e.g. indexing biomedical articles. It is organised as a hierarchy from general categories, such as *Diseases*, *Health Care* and *Geographicals*, to more specific terms in up to twelve levels. Each of the categories has its own tree in the MeSH structure. One term can have more than one place in the hierarchy, for instance the term *Diabetes Mellitus*, which is under the top level category *Diseases*. It can be found in two places in the MeSH tree structure, under *Endocrine System Diseases*, and under *Glucose Metabolism Disorders*, which is an example of *Nutritional and Metabolic Diseases*.

MeSH has been translated to Swedish by the Karolinska Institute Library, Sweden. Swedish MeSH contains translations of 99% of the 26,853 terms in 2013 MeSH. In this work we use MeSH version 2013-01-21, and the Swedish version 2010.

8.2 Methods

The methods are divided into ones related to choice of *sample sets* (section 8.2.1), *normalisations* (section 8.2.2) and *annotation* (section 8.2.3) of search log data to facilitate analyses.

8.2.1 Choice of sample sets

As part of our work we wanted to, if possible, allow for considerations if seeker aspects as *ohälsotal*,[55] *immigration* status, *education* and *age* have an impact on seeker behaviour, but also how they are managed by the health portals. Both 1177.se and Vårdguiden.se offer interfaces and information in different languages, but in our work we focus on the use of the Swedish interface.

[54] www.nlm.nih.gov/mesh

[55] This notion lacks an English translation, but is an estimate of the degree of health related problems in a population, based on number of days with financial support for sick leave or rehabilitation.

We decided to consider the aspects described in table 8.10, aiming at identifying three Stockholm regions and three Swedish counties each containing an average location and two "extremes", called sample sets.

Tables 8.11 and 8.12 present the choices, where the sign "0" indicates a location roughly with an average measure, a "-" a location below and "+" above the average,[56] and in the following sections they are discussed in detail. For instance, considering origin of birth outside Sweden, i.e. whether to be considered an immigrant, the average is around 12% according to SCB 2012 which is close to the proportion for Östergötland, but Jämtland has a proportion which is close to the lowest level and Stockholm is found at the other end of the scale. However, it is important to emphasise that the choices are only based on an overview of the available data. In addition to these, we decided to include the Stockholm region Skarpnäck due to indications of less than 50% of all queries to receive an answer, and Västra Götaland County to allow for comparisons with the Västra Götaland dataset.

8.2.1.1 Age

Age groups were divided into Young, Adult and Elderly, based on available data, tables A.6 and A.7. By table 8.11, Hägersten-Liljeholmen has a population which is close to average for all age groups, Rinkeby-Kista has a more pronounced younger population and Östermalm contains fewer younger and more elderly citizens. Similarly, in table 8.12 for Sweden we find Östergötland in the middle, Stockholm having a younger population and Jämtland an older one in comparison to Östergötland.

8.2.1.2 Education

Considering highest level of eduction, tables A.8 and A.9 present the proportion of the population with given level of highest education, with *gymnasium* 'college' as the divider. For this aspect we find Hägersten-Liljeholmen in the middle, with Rinkeby-Kista having a lower degree of citizens with higher education compared to Östermalm.

Considering Sweden, Östergötland is found in the middle, but Jämtland show a higher degree of college-trained citizens than Östergötland. Stockholm seem to have a smaller proportion of the population with a lower than college education, and a larger population of highly educated citizens.

[56] We denote values "extreme" when they are outside of one standard deviation from the average.

Variable	Dataset	Definition	Division	Source
Age	Stockholm	Percentage of local population in given interval	0–19, 20–64, 65+	Stockholms stad (2013)
	Sweden	Percentage of local population in given interval	0–24, 25–64, 65+	SCB (2014)
Immigration	Stockholm	Percentage of local population not born in Sweden		Stockholms stad (2013)
	Sweden	Percentage of local population not born in Sweden		SCB (2014)
Education	Stockholm	Percentage of local population with given highest education	< college, college, > college	Stockholms stad (2013)
	Sweden	Percentage of local population with given highest education	< college, college, > college	SCB (2014)
Ohälsotal	Stockholm	Ohälsotal of local population in given age group	16–29, 30–44, 45–49, 55–59, 60–64	Försäkringskassan (2012)
	Sweden	Ohälsotal of local population in given age group	16–29, 30–49, 50–59, 60–64	Försäkringskassan (2012)

Table 8.10: Variables used to decide population samples for Stockholm and Sweden.

	Age	Education	Immigration	Ohälsotal
Rinkeby-Kista	+ 0 0	+ 0 -	+	0 + + + + +
Skarpnäck	0 0 -	0 0 -	0	0 0 0 0 0 0
Hägersten-Liljeholmen	0 0 0	0 0 0	0	0 0 0 0 0 0
Östermalm	- 0 +	- + +	0	- - - - -

Table 8.11: Population profiles for selected Stockholm regions.

	Age	Education	Immigration	Ohälsotal
Jämtland	0 + -	0 + 0	-	+ 0 + +
Västra Götaland	0 0 0	0 0 0	0	0 0 + 0
Östergötland	0 0 0	0 0 0	0	0 0 0 0
Stockholm	+ + -	- - +	+	- - - -

Table 8.12: Population profiles for selected Swedish counties.

8.2.1.3 Immigration

Considering proportion of population born outside of Sweden, denoted immigration, tables A.10 and A.11 present the results for the selected locations. For Stockholm, the region Rinkeby-Kista stands out with a larger proportion of immigrants, which is also reflected at a national level. Jämtland is the county with the smallest proportion of immigrants among the selected counties.

8.2.1.4 "Ohälsotal"

Försäkringskassan 'Swedish Social Insurance Agency' defines the notion *ohälsotal* as "antal utbetalda dagar med sjukpenning, rehabiliteringspenning samt sjuk- eller aktivitetsersättning (före år 2003 förtidspension/sjukbidrag) från socialförsäkringen relaterat till antal registrerade försäkrade (befolkningen) i åldrarna 16–64 år. Alla dagar är omräknade till nettodagar, t.ex. två dagar med halv ersättning räknas som en dag." (Försäkringskassan 2011). This notion lacks an English translation, but is an estimate of the degree of health related problems in a population, based on number of days with financial support for sick leave or rehabilitation. In other words, the greater the number the more health related problems there are in the population.

For Stockholm, Rinkeby-Kista is above average, i.e. has more health problems, and Östermalm is found at the other end of the scale, table A.12. However, at a national level Stockholm is "healthier" than average, with Jämtland suffering from more problems than Östergötland, which is a representative of

a county in the middle, table A.13.

8.2.2 Normalisation

The raw data found in the search logs is generally treated as expressed with lower case letters.[57] In addition to this, the only normalisation is trimming of leading and trailing spaces. Not only does our used annotation resources not consider case sensitivity, but also aspects like an inconsistent way of naming concepts makes language processing methods as lemmatisation and stemming in our opinion, less useful without major efforts to normalise the resources which was not in scope of this thesis.

8.2.3 Annotation

We use the UMLS, especially the terminologies Swedish Snomed CT and MeSH (section 8.1.2), as the basis for our efforts to annotate the captured interactions with further semantic details on concepts and relations among them.

The mapping from query terms, or answers, consists of, if not stated otherwise, the query, or its terms, being matched against the UMLS notions called concept names. When possible the mapping of queries was done automatically, using MySQL equality matching, by first trying to map the query to Swedish Snomed CT and then, for queries which could not be matched, trying to map to, in turn, Swedish MeSH and the whole UMLS. The matching, when manual, may involve management of inflections, choice of, in our opinion, adequate concept, etc if so stated, but normally a basic exact matching is the first choice. Thereby, queries have been matched to sequences of sets of concepts to allow further analysis of the search log entries. For instance, the query *feber hosta* 'fever cough' will be mapped to the sequence (C0015967, C0010200) of the corresponding unique concept identifiers named *Fever (finding)* and *Cough (finding)* by the US Snomed CT. When the matching was manual, by the author, the annotations are marked with an asterisk (*).

When mapping answers, only for the Västra Götaland dataset, an answer is given as a hierarchical sequence of terms manually mapped to adequate UMLS concepts and semantic types, see example table 8.8.

However, it is important to remember that our annotation efforts are mainly intended for analyses of our search logs, and not to describe how annotation

[57] The use of upper and lower case letters may make a difference, but due to the inconsistent use in the UMLS, we only consider case sensitivity in analyses of capitalisation and acronyms (section 10.2.1).

may be carried out in a portal setting. This topic is discussed in the sections of chapter 10 on detection of different types of queries and answers, and how this information may be used to decide adequate answers for given queries and contexts. Moreover, in this chapter different problems with mapping methods and resources are discussed.

8.2.4 Analysis

The main types of analyses carried out in this thesis can be divided into statistical ones, methods to induce search games from search logs for analysis and potential decision support, and to identify search scenarios utilising the search logs or games.

8.2.4.1 Search statistics

The search log and annotation data are kept in a MySQL[58] relational database and calculations are made using SQL or in connection with mathematical operations implemented in R.[59] The R environment is also used to generate the graphs found in the thesis.

8.2.4.2 Interval grouping

One of the most challenging properties of the search logs is that it is difficult to divide interactions into groups, where as in the case of sessions for Västra Götaland one interval corresponds to the interactions by one seeker. Moreover, in many cases data is duplicated in unclear ways in the logs. To be able to analyse the data in the light of these challenges, we have decided to divide the interactions into intervals based on the following assumptions:

- no information seeker uses a portal for longer than five minutes
- a seeker posts the same query not more than once per five minute interval

The rationale behind these assumptions follows from an analysis of existing interval information for the Västra Götaland dataset presented in chapter 10 on searches and answers when sessions are given.

[58] http://www.mysql.com/
[59] http://www.r-project.org/

Figure 33 presents the distribution of *session spans*, i.e. time in seconds from first to last registered entry for a given session identifier for Västra Götaland. It shows the majority of sessions take less than a second, hence consist of only one search round. The increase of search rounds around one minute may indicate sessions with more than one round, and we even have sessions with a span over one hour.

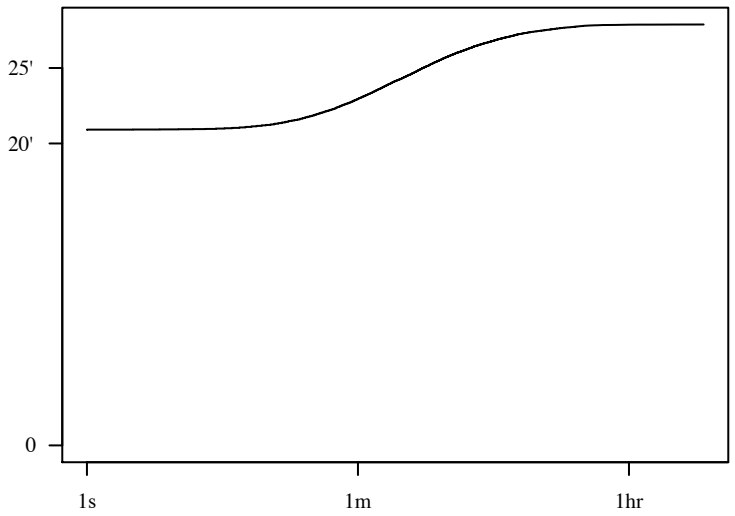


Figure 33: Distribution of session spans for Västra Götaland.

8.2.4.3 Click categories

The search log from Västra Götaland contains click information on the format of internet addresses, see example in table 8.8. An address consists of up to seven parts, divided by the sign "/" and are called *answer levels*, and referenced by their *tree code*. Each of these represent a division of the clicks according to a systematic approach, where each level is manually assigned to a UMLS concept and given a descriptive category, table 8.13. At the fourth level, the last one used in our analysis, the concept is assigned as the category. Consequently, each level may be seen as a facet describing the chosen answer to a query, and the UMLS is used to assign annotations which also make it possible to relate answers to each other, but also to relate answers and queries, utilising the semantic network of the UMLS.

Level 1	Level 2	Level 3	Concept	Semantic type
Inquiry	Advice	Drug	Xerostomia	Sign or Symptom
Inquiry	Disease	Disease	Advice	Health Care Activity
Inquiry	Advice	Group (patient)	Borrelia	Bacterium
Inquiry	Procedure	Procedure (diagnostic)	Advice	Health Care Activity
Inquiry	Theme	Pregnancy	–	–
Facility	Location	–	–	–

Table 8.13: Example of semantic structure of Västra Götaland answers.

8.2.4.4 Search game induction

To induce search games from the search logs, we follow the outlined methods in chapter 6.

8.2.4.5 Identification of search scenarios

Since we do not aim at identifying sophisticated patterns, but mainly ones derivable from an analysis of the syntax and semantic annotation of the queries, as described in chapter 6, we will use basic statistical analysis as described above.

9 QUERIES WITHOUT ANSWERS

When a seeker posts a query to a health portal, her interest is often to obtain answers providing information on, for instance, treatments and health services. However, in some cases the seeker may end up without any answers, due to the portal not being able to interpret her utterance. For instance, if she posts the query *gynokolog*, which is a misspelling of the Swedish word *gynekolog* 'gynaecologist', then the answers depend on the ability of the portal to identify the query to be a misspelling, or on the seeker realising her mistake and correcting her query. Moreover, in some cases the query may be correctly spelt, but the portal not being "aware" of the concept. For instance, since *smittkoppor* 'smallpox' is considered to be extinct, the portal may not include information on this disease, hence it cannot provide any answers to queries related to the disease and the seeker may end up confused and possibly consider the portal to act in an irrational way.

To establish and maintain the seeker's trust in the portal, the portal needs the ability to interpret misspelt queries and handle problems to identify adequate answers. Hence, this chapter will focus on the impact of misspellings and interpretation problems on cooperation, both by examples from selected parts of Stockholm and counties of Sweden. We will also elaborate on the types of deviations which have to be managed by a portal, such as ones where seekers do not realise the reason for the problem, and those where they correct, or change, their utterances as a result of the lack of answers. This part benefits from terminology and research introduced in section 6.5.1 on *error analysis*, i.e. the cause and type of errors, made in second language acquisition (James 1998), and our results show that many of the problems are similar to ones often found when learning a new language like the one of medicine. We have, for instance, examples where terms in the language of medicine are difficult to spell, e.g. *epilepsi* 'epilepsy' and misspellings like *cancer* spelt *canser*. There are also cases where it is possibly unclear to the seeker whether the portal prefers medical terms such as 'colon' to have the Swedish spelling *kolon*. We estimate that in the range of 10% of all incomplete search rounds may result

from problems related to different types of misspellings (section 9.3.1), hence conclude that it is an important area for portal providers to address to maintain seekers' trust, especially since the seekers not knowing their queries to be misspelt view themselves as acting trustworthily.

As will be presented, the interpretation of utterances by existing portals depends not only on the query itself, but also on its context. For instance, if the portal considered the user to be located in Partille (a city in Västra Götaland), the query *akutmottagning* 'emergency unit' would not result in any answers, i.e. an incomplete round in the search game, but for a seeker in Gothenburg it would be interpretable by the portal and result in a complete round with several answers. The possibility of answers to be identified not only in the light of the underlying *vocabulary* of the portal, but also with respect to *context* aspects like seeker location and search time, is elaborated on. We also discuss many of the incomplete rounds, resulting from, in our opinion, a too conservative treatment of context information limiting the answers in a way which may result in distrust and increased problems for portals to achieve their aim of providing health information and support to the public. For instance, as in the case of Partille which does not have an emergency unit, but one is available a few kilometres away in Gothenburg.

In the first section we start by an overview of the problems with cooperation among the actors in a search game with a focus on *incomplete rounds*. Using the samples of the Stockholm and Sweden datasets described in chapter 8, we identify both specific and common problems to be addressed in the following sections. These are divided into ones related to *context dependency* (section 9.2), *misspellings* (section 9.3) and *unknown queries* for the portal (section 9.4), and each area is then discussed in detail to describe current challenges and potential future means to address them. However, we would like to emphasise that our aim is mainly to bring the challenges to the attention of portal providers, since the topic of misspellings and used seeker expressions have, to our knowledge, not been studied for Swedish health portals, and not to present solutions on how to solve them. The importance of this area of research may also be supported by our estimate of up to 5% of the seekers' information needs being unresolved due to portal interpretation problems, and potentially this may seem as irrational behaviour to the information seekers.

9.1 Incomplete search rounds

A search round is *incomplete* whenever the list of answers, provided by the portal as part of its move, is empty. Hence, the seeker posted a query, but did not receive any answers. By table 9.1 we have that, according to our datasets,

incomplete rounds are rather common with more than 10% of all search rounds leading to unanswered search queries. If we also consider that almost 50% of all sessions contain only one round (section 8.1.1), we estimate that 5% of the plays might be incomplete.[60]

	Intervals		**Search rounds**	
	Answers	No answers	Complete	Incomplete
Stockholm	204,233	200,547	5,500,712	1,038,383 (15.9%)
Sweden	109,258	97,529	7,644,566	956,030 (11.1%)
Västra Götaland	27,881	–	66,783	–

Table 9.1: Distribution of intervals and search rounds for the datasets Stockholm, Sweden and Västra Götaland.

It is also worth noticing how incomplete rounds occur over time, with a non-negligible number found during evenings and nights, figure 34, when information seekers might be in more urgent need of adequate support and information. In addition, as will be seen below, a complete round does not mean that the answers were satisfying to the seeker, for instance, answers pointing to health centres open only during daytime when, in the middle of the night, searching symptoms of serious diseases in need of immediate care. Thereby, the collaborative efforts of the seeker and the portal are inadequate, and may impact the seekers' trust in the portal as able to fulfil their information needs during different times of the day.

If we divide the incomplete rounds into groups based on our knowledge of the location of the seeker, tables 9.2 and 9.3, we have that the incomplete rounds range from 10.2% to 57.4% of all rounds depending on the considered location. However, since the datasets Stockholm and Sweden differ in several aspects (chapter 8), we will discuss each set separately. Still, at a local level this indicates a risk that more than one out of ten seekers will not obtain any support for their needs, and based on our choice of studied locations possibly independent of parameters like age, education and language skills. Though, it is worth noticing that Rinkeby-Kista may suffer slightly more than the other considered parts of Stockholm, and is also the one with highest proportion of people with health problems.

[60]The search logs of the health portals contain many utterances leading to incomplete rounds which seem to result from technical problems. The majority of these have been excluded from our analysis, since they are considered to be out of scope of our research. Consequently, the actual risk of ending up with unanswered queries may be substantially greater than presented here.

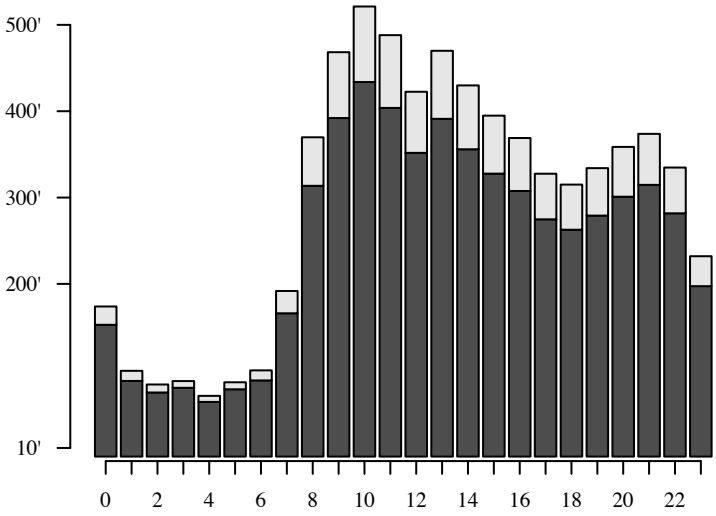


Figure 34: Distribution of complete (black) and incomplete (grey) search rounds for Stockholm over 24 hours.

Location	**Search rounds**	
	Complete / Incomplete	Incomplete
Skarpnäck	87 / 117	57.4%
Rinkeby-Kista	283 / 162	36.4%
Östermalm	853 / 347	28.9%
Hägersten-Liljeholmen	803 / 231	22.3%
Total	2,026 / 857	29.7%

Table 9.2: Distribution of search rounds for selected Stockholm regions during October–November 2010.

Hence, in addition to time-dependent aspects, a portal's interpretation of the seeker's needs, as expressed by the utterances, and its preference relation may have to consider local aspects, based on underlying demographic parameters.

If we study common queries without answers for Stockholm and Sweden, table 9.4, we have that they share many queries leading to incomplete rounds, still being correctly spelt. For instance, health care units such as *akutmottagning* 'emergency unit' and *jourcentral* (emergency unit with limited access) are examples of the types of concepts that often result in incomplete rounds.

Location	Search rounds	
	Complete / Incomplete	Incomplete
Jämtland	91,047 / 14,112	13.4%
Östergötland	453,650 / 51,340	10.2%
Västra Götaland	1,122,365 / 139,453	11.1%
Stockholm	854,266 / 135,201	13.7%
Total	2,521,328 / 340,106	11.9%

Table 9.3: Distribution of search rounds for selected Swedish counties.

Stockholm	**Sweden (Stockholm County)**
akutmottagning	akutmottagning
missbruksvård	vårdcentral
barnavårdscentral	närakut
sesamsös, hudklinik, södersjukhuset ab	kvinnohälsovård
tandvård barn ungdom	barnavårdcentral
bröstmottagning specialistläkarmottagnin	borelia
sex och samlevnad	jourcentral
mottagningssöklänkar	t.ex. mottagning eller typ av vård
psykiatrisk vård	gynekolog
ortopedmottagning sabbatsberg	medicinkliniken
närsjukhuset	

Sweden	**Sweden (unspecified county)**
t.ex. mottagning eller typ av vård	vã¥rdcentral
kvinnohälsovård	borelia
sex och samlevnad	tandvã¥rd
akutmottagning	www.1117se.com/
bb	webbisar
närakuten	frkylningsblsor p tungan
sjukhus	molusker
jourläkarcentral	sår elida
jourcentral	kroniskt trötthetssyndrom
blodcentral	cervikal erosion

Table 9.4: Common queries with no answers for different (parts of) datasets.

Firstly, as discussed in chapter 8, it is possible for the seeker, or by use of portal-induced "cookies", to constrain answers by choosing location and domain of interest, and this may be the reason why queries like *akutmottagning* 'emergency unit' often result in incomplete rounds. That is, if the seeker, possibly unintentionally, has chosen a location, only health units "belonging" to that location are returned as answers. Secondly, some of the queries are rather long and complex, e.g. *sex och samlevnad* 'sex and relations', and may originate from expressions proposed by the portal to the seeker as misleading "feasible" queries. That is, the portal has proposed a query when the seeker start to type, but when she clicks the suggestion it leads to no answers. Thirdly, we also have queries such as *t.ex. mottagning eller typ av vård* 'e.g. health unit or type of care', which seem to be suggestions by the portal to post certain types of queries, but are chosen by the seekers as queries, and these lead to no answers. As seen for Sweden (unspecified county) there may also be problems of a technical nature related to, for instance, character sets. Finally, it is worth noting that the only obvious "misspellings" in table 9.4 are *borelia* which should be spelt *borrelia* and *molusker* which should be *mollusker*.

To summarise, incomplete rounds are rather common and may result from, for instance, contextual *constraints*, misleading query *suggestions*, misunderstood *instructions*, technical *problems* and *misspellings*, but also challenges with preferences, or needs and ways of expression, changing with respect to *time* and *location*.

In our game-theoretic model these findings indicate that the post-query situation in the implemented instances often tend to lead the portal towards problems, in the eyes of the seeker, to decide moves with an ability to satisfy the seeker's needs. In other words, the context part of an utterance tends to have a major impact on both the seeker and portal satisfactions. Moreover, in cases when the portal "promises" improved user satisfaction by proposing queries or giving instructions, in some cases these tend to lead to total dissatisfaction by incomplete rounds, hence an unfruitful collaboration between seeker and portal. This problem is serious, since in these cases the seeker has acted, seen from her own perspective, in a trustful way, being willing to depend on the portal's suggestions. Consequently, with in the range of 5% of all search plays being incomplete, due to for instance contextual impact, misspellings and other communication problems, we will study these in the following sections and elaborate on their importance in the light of a template for portal solutions.

Based on our analysis, to our knowledge the first one considering official Swedish health portals, more than 5% of all information needs may end up unresolved. In the light of the discussion in chapter 2 on the role of internet in today's and future health care, our results may highlight an area in need of further research, as well as discussion and efforts to maintain, and possibly

increase, the trust in the health portals. From a natural language processing perspective, these efforts should, in addition to the used language per se, also consider aspects like time and location, or underlying demographics, of the interactions.

9.2 Impact of context

As discussed in the presentation of the search logs (chapter 8) and the previous section, contextual dimensions, e.g location, used device and domain of interest, not only constrain the user's interest, but also the provided answers. To further study these aspects, we introduce a division of queries based on their *context dependency* and *interpretability* from a portal perspective.

9.2.1 Context dependency and query interpretability

Tables 9.5 and 9.6 present some common weakly context dependent and independent queries. We note that many of the weakly context dependent and independent queries correspond to ones related to health care units, e.g. specific types of units or professions, and ones related to symptoms, diseases and other non-location specific concepts, respectively. If we view these over time, figure 35, searches related to *vårdcentral* 'health centre', are common during evenings and nights, even though these facilities are not open at this time. Moreover, we even notice weak context dependency for this type of queries.

The relation between context dependency and types of queries is addressed in detail in chapter 10, especially in the light of the provided answers and user preferences. However, already by tables 9.5 and 9.6 and the discussion above, we may assume that the seeker expects answers related to advice on what to do to be context dependent to a greater extent than ones related to general information on diseases like chickenpox and fever.

Considering mobile-based searches, i.e. searches logged as having taken place using a mobile device such as a smartphone, table 9.7 presents common queries without answers for mobile in comparison to non-mobile devices. The results might indicate that mobile search problems more often concern diseases and non-mobile search problems health care facilities and professions. However, according to tables A.1 and A.2 there seems to be no substantial difference in the type of information searched for, but as presented in (Eklund 2013a) there seems to be a greater interest among mobile searchers in pregnancy related matters. Hence further research is needed to address questions related to difference in the ability to obtain answers for mobile in comparison

Stockholm	**Sweden**
akutmottagning	vårdcentral
gynekolog	tandvård
vårdcentral	cancer
psykiatrisk öppenvårdsmottagning	barnavårdscentral
ögon	barnmorskemottagning
vattkoppor	hälsocentral
barnmorskemottagning	sjukhus
feber	akutmottagning
sex och samlevnad	psykiatrisk vård
barnavårdscentral	ungdomsmottagning

Table 9.5: Common weakly context dependent queries for Stockholm and Sweden.

Stockholm	**Sweden**
diabetes	bett
sesam, solna	leder
abortmottagning	kost
ortopedkliniken danderyds sjukhus	knä smärta
ortopediska huset johanneshov	hjärtinfarkt symtom kvinnor
mälargården rehab center	akuten
mens	ont i vaden
bvc	finnar
humlegården, rehabilitering	blodgrupper
sommarsol	yrsel och ont i nacken

Table 9.6: Common context independent queries for Stockholm and Sweden.

to non-mobile seekers.

To summarise, according to the search logs, context dependency occurs for both locations and devices, and based on its distribution over time may have a different degree of consequences for the seeker during evenings and nights.

Considering a search game template for future portal solutions, this highlights the impact of not only the query but also its context on the portal moves, and the risk of ending up with incomplete plays whenever the portal does not treat the context information in a sensible way. Thereby, addressing questions regarding the bounded rationality of the portal in the eyes of the seekers should be considered. For instance, assuming the seeker to have the same expectations independent of location and time, e.g. to be provided with available health care units when certain types of utterances are made. If the portal ignores the con-

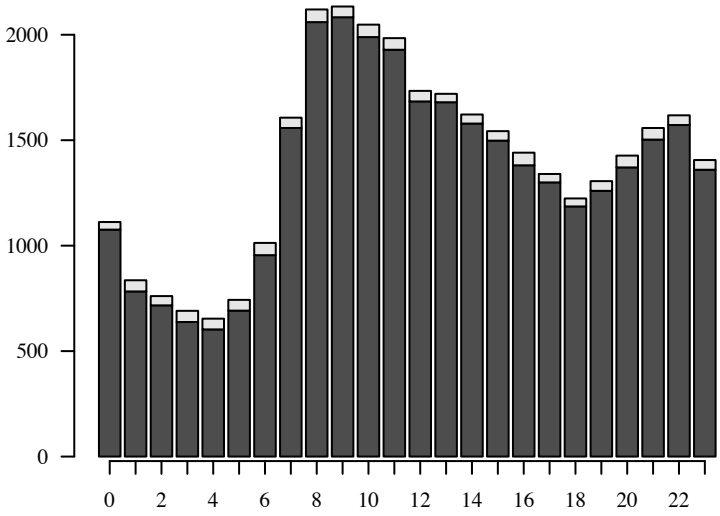


Figure 35: Distribution of answers (black) and no answers (grey) for the query *vårdcentral* 'health centre' over 24 hours for Stockholm.

Mobile	**Non-mobile**
bröstmottagning specialistläkarmottagnin	akutmottagning
sesamsös, hudklinik, södersjukhuset ab	missbruksvård
diareé	barnavårdscentral
akutmottagning	tandvård barn ungdom
hjärtattack	sesamsös, hudklinik, södersjukhuset ab
blodblåsa	sex och samlevnad
kåvepenin	bröstmottagning specialistläkarmottagnin
deprission	psykiatrisk vård
knäskada	mottagningssöklänkar
gynokolog	ortopedmottagning sabbatsberg närsjukhuset

Table 9.7: Common mobile/non-mobile queries without answers for Stockholm.

text in its definition of its preference relation, the seeker will view the portal's behaviour as more, or less, rational depending on her location and time of search, since the probability would be higher for the portal to provide open units close to the seeker during daytime than during evenings and nights. This also impacts the seeker's trust in the portal. She may view its behaviour as unclear and doubt its competence and willingness to support her. The latter may lead to the seeker's trust beliefs being turned into disbeliefs and ultimately

affect her behaviour.

Table 9.8 shows some common partially interpretable queries for Stockholm, where many are location-dependent, e.g. they are related to health care units or professions.

Hägersten-Liljeholmen	**Rinkeby-Kista**
akutmottagning	akutmottagning
barnmorska	psykiatrisk öppenvårdsmottagning
mödravårdcentral	dietist
cellprovtagning	gynmottagning
ögonläkare	health center
ögon	gynekologmottagningen mama mia
sesam city	urolog
ögon specialistläkarmottagning	kbt
cellprov	psykiatrimottagning
barnavårdcentral	gynekologmottagningen betania

Skarpnäck	**Östermalm**
gynekolog	sex och samlevnad
sex och samlevnad	psykiatrisk öppenvårdsmottagning
arbetsterapeut	hudläkare
mödravård	ögonläkare
psykiatri	ögon specialistläkarmottagning
barnmorska	psykiatrisk akutvård vuxen
gynekologmottagning	mödravårdcentral
mödravårdscentral	sesam city
rehab	hemrehab
beroende	gynakut

Table 9.8: Examples of partially interpretable queries with respect to location for the selected Stockholm regions.

By table 9.9 we also have that queries like *akutmottagning* 'emergency unit' are partially interpretable across several datasets. Considering the context dependency discussed above, these results are not surprising. However, an interesting aspect of interpretability is that a query which, for a specific location and time, received no answers can be interpretable when posted at another time, figure 36. Table 9.10 shows some of the queries shifting interpretability over time for the studied Swedish counties, with many being related to health care units. This type of "fluctuating" interpretability implies that the portal's behaviour changes over time. Some variation over time may occur due to new health care units, changed search algorithms etc, but considering the presented

data, the bounded rationality of the portal may be questioned. In other words, if the portal acts in a non-rational way over time in the eyes of the seekers, it may complicate a fruitful cooperation. Consequently, the seeker's trust in the portal may decrease, but being the only official, and by the authorities promoted channel of health information, this could lead to seekers trying to change their behaviour, i.e. their preferences and way of expression, to try to adapt to the portal or, in the worst case, not seeking advice or care as needed. Further on, the changed seeker behaviour may be viewed by the portal as an indication of non-rational seeker actions, but since its underlying search engine tends to be rather fixed not being able to adapt and leading to further gap between the seekers' expectations and the provided support by the portal. Hence, this further stresses the importance of portal solutions behaving rationally, and being trustworthy, in the eyes of the seekers, and the template of these to ensure a sensible treatment of their preferences over time to avoid unwanted changes in seeker behaviour.

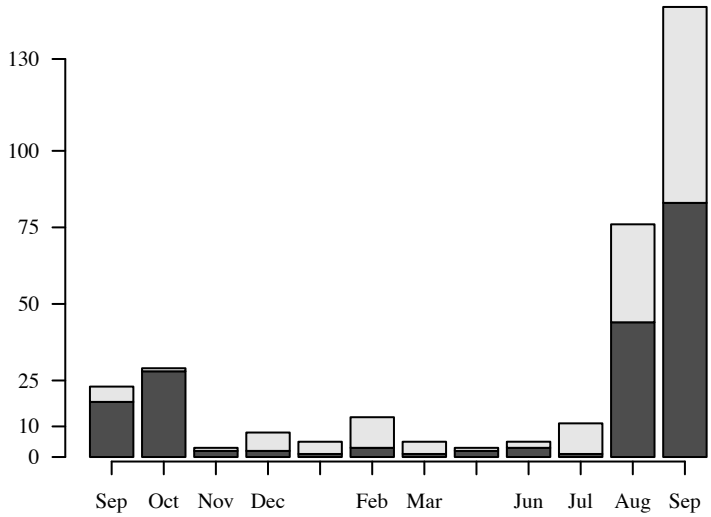


Figure 36: Distribution of the query *gynekologmottagning* 'gynaecologist' over time for Stockholm County, with interpretable (black) and non-interpretable (grey) queries.

To summarise this section, the outcome of a search play is highly dependent on contextual aspects such as the seeker's location, and the interpretability seems to change over time in unclear ways. Moreover, due to differences in seeker behaviour over time, e.g. type of posted queries and use of mobile devices, the outcomes may be less satisfactory during times when seeker needs

Stockholm	**Västra Götaland**
akutmottagning	akutmottagning
vårdcentral	t.ex. mottagning eller typ av vård
närakut	jourcentral
barnavårdcentral	sjukhus
kvinnohälsovård	kvinnohälsovård
jourcentral	psykiatri, vuxna
t.ex. mottagning eller typ av vård	jourläkarcentral
medicinkliniken	gynekologi
gynekolog	vaccinationsverksamhet
blodcentral	gynekologisk akutmottagning

Östergötland	**Jämtland**
t.ex. mottagning eller typ av vård	psykiatrisk vård
mammografi hälsokontroll	närakuten
ombokning	
barnmorskemottagning	sjukhus
kvinnohälsovård	sex och samlevnad
blodcentral	psykiatri
akutmottagning	kvinnohälsovård
psykiatri, vuxna	blodcentral
medicinmottagning	bb
borelia	medicinkliniken
jourläkarcentral	jourcentral

Table 9.9: Examples of partially interpretable queries with respect to location for the selected Swedish counties.

may be more critical, e.g. during nights when only some health care units are open. In the next section we will return to how these aspects may be taken into consideration by a portal to achieve a rational portal behaviour in the eyes of the seeker, and ultimately a trust by the seeker and predictable seeker behaviour allowing natural language processing solutions incorporated in the machinery of portals to achieve better results.

9.2.2 Location- and time-dependency

In the previous section we saw how moves by the studied portals are heavily context dependent, especially considering location and time. However, it is important to stress that context dependency is not a problem as such, but becomes

Stockholm	**Västra Götaland**
gynekologmottagning	psykiatri, vuxna
vuxenpsykiatri	gynekologisk akutmottagning
abortmottagning	kiropraktik
barnmottagning	petekier
psykoterapi	mödravårdcentral
morbus crohn	mag-tarmmottagning
barnläkarmottagning	förlossningsmottagning
petekier	ortopedteknisk verksamhet
arbetsterapeut	hörselvårdsverksamhet
kvinnoklinik	alkoholsjukvård

Östergötland	**Jämtland**
mödrahälsovård	habilitering
naprapat	akutvård
neurologmottagning, medicinska specialistkliniken	logoped
kiropraktor	psykoterapi
hjärtavdelning	abortmottagning
kub test	barnmorska
endokrinologmottagning	ortopedi
beroendemottagning	kiropraktor
jourtandläkare	förlossningen
vuxenpsykiatri	gyn

Table 9.10: Queries with ~50% risk of not receiving answers for the selected Swedish counties.

so when it leads to what the seeker may perceive as irrational portal behaviour, and in the end to distrust of the portal's ability to address the seekers' needs. Moreover, by the nature of the interaction, the means available to the portal, and seeker, to build and maintain trust is to interpret current and past queries in the given contexts as well as possible, and, if needed, drive the portal or seeker behaviour in directions to improve the trust, and consequently provide better answers to the seekers.

The reason for portal moves to be context dependent is to provide the seeker with more accurate answers, hence to better satisfy the seekers. For instance, if the seeker is located in Skarpnäck and as part of her move provides the query *gynekolog* 'gynaecologist' she is probably interested in health care units providing this type of profession nearby and not in an other part of Stockholm.

If the portal's use of the location in its preference relation is "strict", lack of this profession for the given location will result in an empty list of answers, and probably an unhappy seeker. Still, the portal's move was rational in the sense that if there are no gynaecologists in Skarpnäck, the best, and only rational, move is to let the seeker know this by returning no answers. However, to the seeker this may look irrational, since she may not know that this is the case or that the portal applies the constraint in a strict manner. Hence, the portal has to either clarify its answer or weaken the limitations to build seeker trust.

9.2.2.1 Example of rational context dependency

To show that the theory of rationality and preference relations introduced in chapter 4 can be used to elaborate on rational location- and time-dependent preference relations, we will make use of the example above.

As a starting point, we believe location-dependency has to be "gradual", i.e. first promoting local care-givers, but if none are available expanding the search to the neighbourhood of the given location. Thereby, the portal will provide answers which are as good as possible, given the location constraints, and maintaining the seeker's trust. In game-theoretic terminology, the bounded rationality of the portal should promote locally in favour of non-locally constrained answers, but not at the cost of not being able to make a move. Viewing this from an information seeker perspective, it means that the portal assumes the seeker to prefer answers referencing local care givers in favour of more distant ones, but the latter are better than no care givers at all. That this type of seeker behaviour is rational in some cases is proven if the implemented preference relation is complete and transitive.

Given any two alternative answers α and β referring to health care units a and b providing a given service 24/7, let the former answer be preferred, $\alpha \succeq \beta$, if the distance from the seeker to a is less than or equal to the distance to b. Then this relation is clearly both complete and transitive, hence rational. If we extend the relation with a unit providing a given service to be better than one not providing it, and if none of them does so, the seeker is indifferent. Then this extended relation is also complete and transitive, thereby the proposed method would satisfy the requirements of a rational relation taking into account location as a basis for preference.

Similar problems may also occur if a seeker searches for care givers at certain times of the day. For instance, searching for *jourcentral* (emergency unit with limited access) in the middle of the night may result in answers, but all of them are open only early evenings and weekends. Hence, as well as utilising a less strict filtering for location-dependency, queries where time is an

important aspect have to be taken into consideration. In other words the portal may assume the seeker to prefer care givers who are open, given the time of search, in favour of closed ones, but the latter in favour of no answers. If we assume answer α and β to represent answers referring to two health care units a and b at the same distance from a seeker, and we let $\alpha \succeq \beta$ whenever a is open longer than b with respect to the current point in time, then the relation is both complete and transitive along the same line of reasoning as above. Moreover, if we extend the relation with the assumption that any answer referring to an open unit is better than a closed one, the extended relation will also be rational.

If we now want to define a rational preference relation among answers considering both time and location, we may let answer α (referring to unit a) be strictly preferred over β (referring to unit b) with respect to a location c at a time t, where $\delta(x)$ is the distance from c to x and $\Delta(x)$ is the remaining open hours with respect to t for $x \in \{a, b\}$, as follows:

$$\alpha \succ \beta \text{ if } \begin{cases} \delta(a) < \delta(b) & \Delta(a), \Delta(b) > 0 \\ \delta(a) \geq \delta(b) & \Delta(a) > 0; \Delta(b) = 0 \end{cases}$$

If the prerequisites do not hold, the portal is indifferent with respect to the answers. First of all the relation is complete, since for two answers the portal will prefer the closest one which is open, and if none is open the portal is indifferent. If we assume $\alpha \succ \beta$ and $\beta \succ \gamma$, then clearly we have that $\alpha \succ \gamma$. Hence the relation is transitive, and we may, in theory, implement a rational portal preference relation among answers which considers both location- and time-dependency. By our discussion on rationality and trust in chapter 4, this relation would also be the basis for trust in the eyes of an information seeker since it shows a trustworthy behaviour.

9.2.2.2 Summary

On the one hand context dependency, e.g. location- and time-dependency, is desirable, and realisable maintaining portal rationality, since it emphasises the provision of answers adjusted to the seeker's location and the time of search. In a search game setting, this implies that in any instance of a search game context dependency, especially considering location and time, is desirable, thereby to be part of a search game template. However, in some cases, for instance when the seeker wants to know more about a disease, neither of these location and time adjustments are expected, or considered rational, by the seeker. Hence, the portal must, to the best of its abilities, try to detect when to consider these types of aspects, and if the benefits would be justifiable in the light of the needs of the seekers and aims of the portal providers.

These questions are the topics of the rest of this section, and we begin by studying the ability to detect location- and time-dependency and the potential benefits based on analysis of the search logs. However, context dependency is only one aspect to consider, and we then turn our interest towards the cases when the seeker provides non-interpretable utterances, independent of context.

9.2.3 Detection of location- and time-dependency

Detection of location-dependent queries is not as simple as just validating potential occurrences of contextual constraints referencing locations or cases when the domain of interest to the seeker has been set to health care units. The main reason is that by the interface of the portal a location is almost always set as a default based on "cookies" or identification of seeker internet address, probably with the consequences unknown to the seeker. Hence, choice of location may almost be seen as a query independent decision.

One approach to address the challenge of identifying potential location-dependency, without justifying it based on locations provided as part of the context, is by semantic annotations using sources like the Swedish Snomed CT's and MeSH's categorisation of medical concepts, to try to identify types of queries more probable to be related to specific locations. Possible candidates are concepts regarding *Health Care Related Organizations* and *Health Care Activities*. As examples of these we find queries like *vårdcentral* 'health centre' and *tandvård* 'dental care', table 9.11.[61] Hence, being able to tag utterances with medical concepts the portal will obtain guidance on whether to consider location specific information when deciding its move. However, some commonly used queries related to health care organisations and activities are not possible to easily identify by this type of annotation. For instance, compound queries with references to *mottagning* 'surgery' and *central* 'centre' are among these, even when they are clearly related to locations. Studying Stockholm County, we have that, for instance, all queries which referenced *jourcentral* (emergency unit with limited access) ended up as incomplete search rounds.

It is not only queries explicitly related to organisations or professions which may indicate an interest in location specific information. For instance, by an analysis of the Västra Götaland dataset, we have that in almost 23% of the cases when a seeker posted the query *stroke* she has chosen answers related to *SU Stroke Forum* specialised in stroke rehabilitation. Hence, queries which

[61] When concepts or semantic types are presented with an asterisk (*) it indicates the, in our opinion, relevant associated concept or type, but where the query is not found in the considered resources.

may indicate post-query situations of an interest in rehabilitation may point towards an interest in location constrained answers. The relation between types of queries and answers are further discussed in chapter 10.

We estimate that 20% of all interactions are location-dependent, of which 75% are explicitly chosen health care units, and by basic approaches detectable by the portal, see chapter 10 for details. Expressed in terms of our theoretical framework, if a search game instance does not try to detect, even when this is a challenge, and account for location-dependency there is a non-neglectable risk of incomplete search rounds and unsatisfied information seekers.

As mentioned above, location-dependency is not only related to queries and given location information, but also to the time of seekers' moves. For instance, a seeker in Partille posting the query *bröstsmärtor* 'chest pain' in the middle of a Friday night would probably not benefit from information about the local *vårdcentral* 'health centre' which is open on weekdays. Instead she would probably want to know that she should visit the *akutmottagning* 'emergency unit' in Gothenburg. Hence, in addition to tagging queries as location-dependent, some of them would benefit from indicators of time-dependency. Time-dependency is obviously related to health centres, which have certain opening hours, but also to types of queries and expected answers by the seekers. For instance, if a seeker searches with the query *fästing* 'tick' in February she might be more interested in prevention than treatment information.

To summarise the discussion on location- and time-dependency, these properties are desirable, but have to be managed in a sensible way to ensure rationality and trust.

Query	Concept	Semantic type
vårdcentral	Health center	Health Care Related Organization
barnavårdscentral	Maternal-Child Health Centers*	Health Care Related Organization*
barnmorskemottagning	Antenatal clinic*	Health Care Related Organization*
akutmottagning	Accident and Emergency department	Health Care Related Organization
tandvård	Dental Procedures	Health Care Activity
hälsocentral	Health center*	Health Care Related Organization*
diabetes	Diabetes	Disease or Syndrome
ungdomsmottagning	–	Health Care Related Organization*
feber	Fever	Finding
feber hos barn	Fever*; Child*	Finding*; Age Group*
gynekolog	Gynecologist	Professional or Occupational Group
sjukgymnast	Physical Therapist	Professional or Occupational Group
sex och samlevnad	–	–
säsongsinfluensan vaccination	Influenza vaccination*	Therapeutic or Preventive Procedure*
psykiatri	Psychiatry Specialty	Biomedical Occupation or Discipline
mödravård	Prenatal care	Health Care Activity
jourcentral	–	Health Care Related Organization*

Table 9.11: Common queries for Sweden, Stockholm and Västra Götaland and their semantic types.

9.3 Misspelt utterances

Even though, as elaborated on in the previous section, context impacts the search play outcome, information seekers do post queries which the portal is unable to interpret independent of chosen context, i.e. queries always leading to incomplete search rounds. For instance, the query *borelia* is often used by information seekers to refer to the disease borrelia. Hence, in this case the portal has to deal with a "traditional" misspelling. In some cases the seekers use queries like *husdoktor* 'housedoctor', which is a correct Swedish compound, but the used term in health settings is *husläkare* 'general practitioner'. This is another type of challenge to be addressed by the portal to try to avoid incomplete search rounds. In these cases, it is important to remember that the seeker most probably considers herself to have acted rationally, not knowing that the queries are misspelt, or as in the case of Google, known to probably be misspelt but corrected automatically by the search engine. Even more important is that even if the portal were to return an empty list of answers, it may not be clear to the seeker what caused this. Tables 9.12 and 9.13 present an estimate of the proportions of non-interpretable queries for the selected Stockholm regions and Swedish counties. Hence, the number of non-interpretable queries seems to be greater than the number of fully interpretable ones for Stockholm, but for Sweden substantially fewer than the fully interpretable ones.

	Queries		
	Non-interpretable	Partially interpretable	Interpretable
Hägersten-Liljeholmen	63 (6.1%)	939 (90.8%)	32 (3.1%)
Östermalm	74 (6.2%)	1051 (87.6%)	75 (6.3%)
Rinkeby-Kista	41 (9.2%)	372 (83.6%)	32 (7.2%)
Skarpnäck	10 (4.9%)	190 (93.1%)	4 (2.0%)

Table 9.12: Distribution of queries for the selected Stockholm regions during October–November 2010.

	Queries		
	Non-interpretable	Partially interpretable	Interpretable
Stockholm	102,004 (10.4%)	654,956 (66.8%)	223,867 (22.8%)
Västra Götaland	111,409 (9.0%)	863,669 (69.7%)	263,872 (21.3%)
Östergötland	35,655 (7.4%)	358,818 (73.0%)	86,869 (18.0%)
Jämtland	5,991 (5.9%)	80,901 (80.2%)	13,946 (13.8%)

Table 9.13: Distribution of queries for the selected Swedish counties.

Moreover, if we compare the proportions of incomplete search rounds with respect to the total number of search rounds, the Rinkeby-Kista region stands out with almost 50% more search rounds without answers, in comparison to other regions. This aspect is also found for Stockholm as a county in Sweden, and considering these they share a substantially larger population of immigrants, for whom Swedish is not their first language. Hence, it raises the question if non-interpretability might be related to Swedish skills, but also why over 5% of all Östermalm queries are non-interpretable, in spite of a population with a high degree of education and very few immigrants.

From a search game perspective, a misspelt query corresponds to an "incorrect" move in the sense that if the portal would interpret it literally, it would result in an empty answer list. However, the seeker may not be aware that the query is misspelt, or even be able to correct the misspelling if being made aware of it. Moreover, by the official role of portals like 1177.se and vardguiden.se, they cannot ignore trying to understand the intentions of the seekers and answer, even in these cases. However, the risk of presenting a "wrong" answer, in the eyes of the seeker, increases in these cases, thereby also the risk of a distrust and changed seeker behaviour. Hence, the treatment of misspellings, and unknown queries, are both very important and difficult, and in our opinion, the first step is trying to characterise and understand the causes of the misspellings.

The questions on the cause and characterisation of non-interpretable queries leading to incomplete plays in the search game instance of Vårdguiden and 1177 are studied in the following sections where focus is on linguistic deviations and unknown queries with respect to the vocabularies of the portals, based on the theory presented in section 6.5.1.

9.3.1 Spelling and vocabularies

By the analysis in section 9.1, we have that more than 5% of the plays in the considered Stockholm regions and Sweden counties are non-interpretable and may result from misspellings or the use of words not known to the portals.

Misspellings are to be defined with respect to a vocabulary and in our case we let the, to us unknown, indexes of the considered portals at the time of capturing of the search logs make up their vocabularies. However, deciding the seekers' vocabulary is much more difficult, but a simple approach would be to let the seekers' vocabulary consist of all posted queries. However, thereby no queries would be misspellings with respect to the vocabulary, even in cases where a seeker clearly changes her query in a following search round. Hence, if we treat the set of queries, excluding queries obviously being portal generated

technical strings of letters, we need a way to measure type and amount of queries which would not be considered correctly spelt by manual inspection of a medical expert. In our case, we use the measures presented in table 9.14 in addition to manual inspection.

Misspelling	**Measure**
Punctuation	Split: non-interpretable query of length two is found as an interpretable query of length one in the same interval.
	Fusion: non-interpretable query of length one is found as an interpretable query of length two in the same interval.
Typographic	Spatial: non-interpretable query is found as an interpretable query with the same string length, and with Levenshtein distance one between the queries.
	Temporal: non-interpretable query is found as an interpretable query with the same string length, and with Levenshtein distance two between the queries.
Dyslexic	Non-interpretable query is found as an interpretable query with the same string length, and with Levenshtein distance one between the queries. Manual inspection identify candidates.
Confusibles	Not measured
Mispronunciation	Non-interpretable query is found as an interpretable query with the same string length, and with Levenshtein distance one between the queries. Manual inspection identify candidates.
Written misencodings	Non-interpretable query is found as an interpretable query with the same string length, and with Levenshtein distance one between the queries. Manual inspection identify candidates.

Table 9.14: Measures to estimate first order misspellings.

9.3.2 Substance

The discussion on vocabularies and the topics of unknown words and misspellings highlights the importance of defining the relevant types of spelling challenges to allow a concerted discussion on their impact and potential management. Moreover, since we noticed in the previous section that the proportion of non-interpretable queries might be greater for regions and counties with more non-native Swedish speakers, a division of spelling changes will facilitate a discussion on the origin of the problems and how to address them. To study the different types of "misspellings" found in the search logs we make use of James' (1998) presentation of error analysis as outlined in section 6.5.1.

As seen for Stockholm County in table 9.15 considering common deviations, many of them are possibly second-order mistakes or errors. For instance, 12.7% of the search rounds in the Sweden dataset referring to the term *borrelia* are misspelt as *borelia*. Considering Stockholm County, we have that in the range of 10% of all non-interpretable queries might be due to spelling problems (see section 6.5.2).[62] If we try to further describe the occurrence of misspellings in Stockholm County, based on the measures in table 9.14, we obtain examples of common misspellings. The first order deviations are estimated as misspelt queries in an interval with a correctly spelt version of that query in the same interval, and in the case none is found the query is considered a second order deviation.

Query		**Search rounds**
Non-interpretable	Interpretable	Incomplete / Complete
borelia	borrelia	313 (18.4%) / 1386 (81.6%)
molusker	mollusker	114 (16.0%) / 599 (84.0%)
lpk	–	100 / –
kontrastvtska	kontrastvätska	87 (91.6%) / 8 (8.4%)
dyspne	–	73 / –
öroninflamation	öroninflammation	71 (10.4%) / 610 (89.6%)
nidblödning	–	66 / –
canser	cancer	66 (5.5%) / 1144 (94.5%)
ichias	ischias	65 (7.6%) / 794 (92.4%)
virusprickar	–	62 / –

Table 9.15: Common non-interpretable queries and their occurrences in Stockholm County.

In the case of punctuation problems, table 9.16 presents some common split deviations found in the county of Stockholm. It is worth noticing that it is quite common that seekers are unsure whether queries referring to a body part and a problem with it, e.g. *collum fraktur* 'collum fracture', should be expressed as one or two words. Other examples of splits are when medical terms are divided into parts, possibly thought to have semantics on their own. Fusions seem to be less common, but it is interesting that a correct term like *smittkoppor* 'smallpox' does not result in any answers, and the query *smitt koppor* to be found in the same interval, hence a split deviation which is interpretable by the portal. Another example is when the term *alkoholförgiftning* 'alcohol poisoning' does not result in any answers, the seekers try the interpretable query *alkohol*

[62] An analysis of the use of the Swedish National Taxboard estimated in the range of 10% of the queries to be "misspelt" or erroneous (Dalianis 2002).

förgiftning. The findings related to split and fusion deviations highlight the importance of a portal to be able to interpret compound terms as one term and as a pair of terms whenever both make sense. This is even more important for medical terms, where it may be unclear how the term should be expressed and both versions seem acceptable to the seeker.[63]

Query	
Non-interpretable	Interpretable
collum fraktur	collumfraktur
border line	borderline
frontal loben	frontalloben
fertilitets mottagning	fertilitetsmottagning
pace maker	pacemaker
lycko piller	lyckopiller
adams äpplet	adamsäpplet
arterio sclerosis	arteriosclerosis

Table 9.16: Examples of split deviations for Stockholm County.

Considering typographic problems found in Stockholm County, table 9.17 presents examples of spatial and temporal deviations, and as expected they are mainly first order deviations where the seeker realises her typographic misspelling and corrects it when the first query did not result in any answers. However, there are still in the range of 10–30% of the cases where this does not seem to be the case and where the deviations should possibly be considered second order. For instance, the misspelling *atsma* of *astma* is in almost half of the cases not corrected.

Moving on to dyslexic problems, table 9.18, we find a more important type of deviation, with a, to us, unexpected number of queries related to common problems like *cancer* and *skabb* 'scabies' being misspelt with respect to the use of similar sounding letters. For instance, the disease cancer is referenced in 13,741 search rounds for Stockholm County, but in 3.8% of the cases it is misspelt as *canser*, *kanser* or *kancer*. In some cases, e.g. when using drug names or uncommon medical terms, we also find this pattern of deviation, but in these cases it may be less unexpected. This type of deviations highlight not only the problem for information seekers to know the spelling of medical terms, but also that rather common terms like cancer are very often misspelt. Thereby we conclude that a portal should try to provide support for this type of problems in its interpreters. Since dyslexic deviations concern cases where one

[63] A study of 128 highly frequent Swedish compound queries with no answers was in (Dalianis 2005) shown to be interpretable in over 60% of the cases if they were decompounded.

letter has been replaced with another with similar sound, this type of problems should be quite easy to address.

Query		**Query**	
Non-interpretable	Interpretable	Non-interpretable	Interpretable
huvudlösa	huvudlöss	atsma	astma
örininflammation	öroninflammation	hjärtinfrakt	hjärtinfarkt
brännsksda	brännskada	lugncancer	lungcancer
ont i halseb	ont i halsen	kjesarsnitt	kejsarsnitt
gallsteb	gallsten	infleunsa	influensa
paykolog	psykolog	infulensa	influensa
blodtryvk	blodtryck	hlasfluss	halsfluss
vårscentral	vårdcentral	tramvred	tarmvred
njutsten	njursten	häjrta	hjärta
nädblod	näsblod	gyenkolog	gynekolog

Table 9.17: Examples of spatial and temporal deviations for Stockholm County.

Query	
Non-interpretable	Interpretable
canser	cancer
systa	cysta
scabb	skabb
brikanyl	bricanyl
kampylobakter	campylobakter

Table 9.18: Examples of dyslexic deviations for Stockholm County.

As in the case of dyslexic problems, mispronunciations and deviations due to written encodings are more difficult to identify, but table 9.19 presents some examples. They show that seekers often have problems spelling medical terms due to their unfamiliarity with the terms and trying to spell it as it sounds "right". This often means choosing among pairs of letters such as e–i, o–e, b–p and u–o. However, one of the terms which seems difficult to spell is *gynekolog* 'gynaecologist' and for Stockholm County there are several different variations of misspellings, e.g. *gynokolog* and *gynekolg*. Still, in the majority of the cases it is one or two letters which have been replaced and, as in the case of dyslexic deviations, a portal interpreter utilising a basic distance measure among words should be able to improve the management of this type of problems.

It may be difficult to differentiate between so called written encoding problems and the ones discussed in this section, but if we just look at some common

Query	
Non-interpretable	Interpretable
epelepsi	epilepsi
gynokolog	gynekolog
deprission	depression
asbergers	aspergers
köttelfeber	körtelfeber
urulog	urolog
sarmonella	salmonella
celebral pares	cerebral pares

Table 9.19: Examples of mispronunciations for Stockholm County.

problems not addressed above we find deviations related to Swedish versus the Latin spelling often used for medical terms. For instance, the body part *colon* is normally spelt *kolon* in Swedish, but in a third of the cases for Stockholm County it is spelt *colon*. These problems are found for many other terms with similar spelling differences between Swedish and Latin, and in this case it raises the question if common Swedish terms in the portal vocabulary should be complemented with their Latin counterparts to reduce the number of deviations due to writer encodings. Another interesting example which is more difficult to explain is the spelling of *inflammation* as *inflamation* in 3.6% of the cases.

In the analysis of Stockholm County with respect to misspellings, we saw that different types of misspellings are common and their origin is in most cases an *unfamiliarity* with medical terms, both with respect to their spelling and how to express compounds referring to different types of care providers or problems related to specific body parts. Since we estimate that in the range of 10% of all non-interpretable queries are caused by misspellings it is a problem which cannot be neglected by the portal. However, as alluded to, many of these may be addressed by portal vocabularies recognising both traditional Latin and common Swedish spelling of medical terms, but also interpretation sensibility of basic dyslexic and mispronunciation deviations. Moreover, the portal must address the challenges of medical terms which are expressible both as a compound term and as a sequence of terms, and recognise both of them. Our analysis shows that today this is not managed in a well-defined way, resulting in a portal behaviour which may seem irrational to the seekers.

From a search game perspective, misspellings will result in incomplete plays if not managed by the portal. Based on our estimates, in the range of 5% of all plays will result in inability for the portal to make a sensible move as

a reaction to the seeker's move. In other words, there is a risk of second order misspellings to result in seekers considering the portal irrational if it does not manage to "correct" the misspelling.

9.3.3 Text

The study of misspellings was part of a substance analysis of the queries and we now turn our interest to text level. The most important type of formal deviations are misformations, i.e. words neither belonging to the seeker nor portal vocabulary. These, including queries belonging to the seeker but not the portal vocabulary, are studied in section 9.4. Among the semantic deviations we find the sense problems, and these are discussed in section 10.2.2, but it is worth mentioning that our analysis reveals problems with synonyms where one query has answers and its synonyms do not, for instance *aneurysm* is interpretable, but its synonym *artärbråck* is not. Another example is *baksmälla* 'hangover' which is interpretable by the portal, but not its hypernym alkoholförgiftning 'alcoholic intoxication'.

A sensible portal treatment of misformations and sense (deviations) is crucial to maintain seekers' trust, especially considering synonyms and hypernyms to ensure a rational management.

9.3.4 Discourse

A basic approach to estimate the occurrence of unrelated terms, or topic coherence problems, is to study the existence of non-interpretable queries containing two terms where the terms together are not mappable to the UMLS, but they are if treated as unrelated terms. In our study we find examples like *ödem sköldkörtel* 'oedema thyroid gland' and *åderbråck varicier* 'varicose veins varices', where the second example is particularly interesting since the two query terms are synonyms.

Pragmatic deviations can be partly estimated as the number of synonyms where some queries result in answers and some do not. Hence, one could claim that the seeker made a pragmatic deviation according to the portal. Using the UMLS, we see that for Västra Götaland we have 2,502 sets of synonyms, of which 896 (36%) contain both non-interpretable and interpretable queries.

9.3.5 Summary

Based on James's (1998) descriptions of different types of linguistic problems, we have that misspellings are a problem, especially dyslexic and proper ones possibly resulting from information seekers' unfamiliarity with the spelling of medical terms. However, we also find many examples of medical versus "common" spelling of terms, like *colon* versus *kolon*, and correct use of compounds and splits which are not managed by the portal vocabulary. Considering textual deviations, many common compound utterances , e.g. *mödravårdscentral* 'antenatal clinic', result in distortions reflecting, for instance, the semantics a health centre for mothers (to be) in comparison to a centre for antenatal care. Still, this type of problems are not as major from a seeker perspective as the ones at substance level.

Independent of the type of deviations, many of them seem to be first order, that is self-corrected by the seeker when she did not receive any answers in her first attempt. However, the problems with the use of medical versus common language and the role of compounds and splits are examples which tend to be of second order, consequently being at risk of leading not only to incomplete search rounds, but also to incomplete games.

As discussed, there are, in our opinion, a handful of rather simple approaches to improve the collaboration between the seeker and portal, hence achieving increased seeker trust in the portal, by including both medical and common spelling of medical terms in the portal vocabulary together with improved support for compounds and splits. In addition to vocabulary extensions, we also propose improved management of basic dyslexic and proper misspellings based on, for instance, Levenshtein distance measures to detect and correct this type of problems. As discussed in the case of text and discourse deviations, these are more difficult to analyse, and probably for a portal to manage. However, even in these cases, basic use of annotation resources and distance measures may provide both insights and a basis for portal methods.

9.4 Unknown queries

In the previous section we studied the occurrences of misspellings and textual deviations leading to non-interpretable queries and incomplete search rounds. However, based on the data for Stockholm County, we estimate that 38% of the queries are non-interpretable and if we exclude the estimated 10% misspellings of non-interpretables, we have in the range of 90% of the non-interpretable queries unaccounted for. Moreover, as will be discussed in chapter 10, only 6% of these are found in medical terminologies like the Swedish Snomed CT

and MeSH. Hence, approximately 32% of all queries are non-interpretables as result of being unknown with respect to the studied health portals' vocabulary and not mappable to any terminology. Thereby we have that with the considered terminologies we will have a substantial number of search rounds being incomplete and resulting in dissatisfied information seekers.

In this section we will study some types of unknown queries with respect to today's portal vocabulary, but also in the light of the results discussed in chapter 10 on extensions of vocabularies try to improve the interpretations of seeker queries.

By table 9.20 we have that among non-interpretable queries we find misspellings, as well as well-defined medical terms, probably expected by the seekers to be found among terms recognised by a health information portal such as 1177 Vårdguiden. Among the terms we find, for instance, *smittkoppor* 'smallpox' possibly missing since the disease is considered extinct and thereby not in the scope of a health information portal. However, we also found queries like *borelia*, which is a very common misspelling, possibly even a second order deviation in many cases.

Spelling	**Medical terms**
borelia	dyspne
molusker	urosepsis
öroninflamation	hypoxi
ichias	smittkoppor
brock	laryngit
canser	
sharlakansfeber	

Table 9.20: Examples of non-interpretable queries for Sweden.

If we focus on the medical terms which would be expected to be members of a portal vocabulary, some can be mapped directly to the UMLS, table 9.21.

By UMLS, we are also able to identify synonyms where only some are interpretable, table 9.22. Similar problems are found among other semantically related terms, table 9.23. Table 9.21 also shows that being mappable to the UMLS does not mean that the mapping has to be correct. For instance, the queries *brock* and *korallen* are mappable, but in the first case the query is mapped to Brock's syndrome, when probably the intended query was *bråck* 'hernia' (possibly with its older spelling *brock*). In the second case the intended query was probably the health centre Korallen. At the same time tables 9.24 and 9.25 highlight that many common queries are not mappable to the UMLS, especially those which are related to health care management or local aspects,

Query	Search rounds	Concept	Semantic type
hemoroider	4,620	Hemorrhoids	Acquired Abnormality
äggstocksinflammation	3,155	Oophoritis	Disease or Syndrome
smittkoppor	2,716	Smallpox	Disease or Syndrome
brock (x)	2,114	Middle Lobe Syndrome	Disease or Syndrome
korallen (x)	1,859	Coral	Eukaryote
tinitus	1,569	Tinnitus	Finding
alkoholförgiftning	1,460	Alcohol Intoxication	Mental or Behavioral Dysfunction
dyspne	1,337	Dyspnea	Sign or Symptom
psykopati	1,298	Antisocial Personality Disorder	Mental or Behavioral Dysfunction
könsstympning	1,105	Female genital cutting	Finding
laryngit	1,065	Laryngitis	Disease or Syndrome
spottkörtelinflammation	985	Sialadenitis	Disease or Syndrome
djurbett	935	Animal bite	Injury or Poisoning
narcissism	905	Narcissism	Mental or Behavioral Dysfunction
ascites	857	Ascites	Finding

Table 9.21: Common non-interpretable queries which can be mapped to a UMLS concept for Västra Götaland County. Possibly incorrect mappings are marked (x).

poorly covered by the UMLS. However, even terms like *klamydia* 'chlamydia', *magsjuka* 'stomach flu' and *bröstcancer* 'breast cancer' are not recognised by the UMLS. Moreover, the occurrence of non-mappable queries seems not to be a specific mobile problem, i.e. herein called 6-queries by requiring more than five answer gists to be considered before an interesting on is found (see section 10.1), thereby it is independent of the portal preference relation.

Query	
Interpretable	Non-interpretable
ångestsyndrom	ångeststörningar
kärlkramp	angina pektoris
angina	
angina pectoris	
antisocial	antisocialt beteende
aneurysm	artärbråck

Table 9.22: Examples of (partially) interpretable queries with non-interpretable synonym for Västra Götaland County.

Interpretable hyponym	**Non-interpretable query**	**Interpretable hypernym**
baksmälla	alkoholförgiftning	
alkoholism		
alkoholberoende		
carcinoid	adenokarcinom	cancer

Table 9.23: Examples of non-interpretable queries with (partially) interpretable hypo-/hypernym for Västra Götaland County.

To summarise, by the UMLS it is clear that the existing portals' vocabularies are inconsistent with respect to their treatment of semantically related query terms such as synonyms. Moreover, they clearly lack commonly used terms expected to be recognised by an official health information portal. However, extending a portal's vocabulary by, for instance, parts of the UMLS would not solve all problems. To understand the breadth of the problem it is important to be aware that neither Snomed CT nor MeSH used as Swedish resources in the UMLS were created to function as look-up databases. The purpose was more to provide a source to promote the use of well-defined concepts in medicine, independent of the used language. Hence, substantial efforts are needed to further develop these resources with adequate terms, phrases and abbreviations used in daily activities both in health care and by laymen. This topic is further discussed in chapter 10, where we see that certain types of queries always

imply post-query situations with expectations for given types of answers. Consequently, the problems of UMLS coverage may thereby be partly managed by improved use of query–answer information.

Query	**Utterances**	**Concept**	**Semantic type**
mina vårdkontakter	1.2%	Patient Care Management	Health Care Activity
egenremiss	0.8%	Self-referral	Health Care Activity
sjukresor	0.7%	Benefits, entitlements and rights	Intellectual Product
urinvägsinfektion	0.6%	Urinary tract infection	Disease or Syndrome
vårdgaranti	0.6%	Benefits, entitlements and rights	Intellectual Product
byta vårdcentral	0.5%	Patient Care Management	Health Care Activity
webbisar	0.5%	–	–
vårdval	0.4%	Benefits, entitlements and rights	Intellectual Product
klamydia	0.4%	Chlamydia Infections	Disease or Syndrome
magsjuka	0.3%	Infectious gastroenteritis	Disease or Syndrome

Table 9.24: Common queries not automatically mapped to UMLS concepts for Västra Götaland.

Query	Utterances	Concept	Semantic type
mina vårdkontakter	6.4%	Patient Care Management	Health Care Activity
egenremiss	4.4%	Self-referral	Health Care Activity
vårdgaranti	3.0%	Benefits, entitlements and rights	Intellectual Product
bröstcancer	0.8%	Malignant neoplasm of breast	Neoplastic Process
svininfluensa	0.8%	Influenza due to Influenza A virus subtype H1N1	Disease or Syndrome
vaccinering	0.8%	Vaccination	Therapeutic or Preventive Procedure
cellprov	0.7%	Pap smear	Diagnostic Procedure
torslanda	0.5%	–	Geographic Area
lista	0.5%	Patient Care Management	Health Care Activity
journaler	0.5%	Medical records	Intellectual Product

Table 9.25: Common 6-queries not automatically mapped to UMLS concepts for Västra Götaland.

9.5 Properties of a trustworthy portal II

In this chapter the focus has been to analyse queries without answers, and the importance of this group of queries is supported by our findings that in the range of 5% of the information seekers' needs may not be addressed in a proper way by the existing health information portals due to interpretation problems. The consequences of this is further emphasised by the perspective that many of the portals' moves are possibly viewed as irrational by the seekers, and may even impact the trust in the portal's "competence" and "willingness" to support the public. Moreover, to our knowledge, this analysis of 1177.se and vardguiden.se is the first attempt of an in-depth description of the existing use of the portals, and is aimed at highlighting both challenges and opportunities for the portal providers.

One of the most important aspects of the existing portal solutions is their treatment of location and time information as part of the context of a query. Constraining the portal's answers to ones related to a certain location is often in the interest of information seekers, but it could also result in no answers to queries which in the eyes of a seeker should result in several. Hence, location, and time, are aspects which are very important for portal providers to treat in a sensible way. This challenge was already discussed in section 6.7 as the first principle to satisfy by a trustworthy portal. As seen in the case of the studied portals, this is a major challenge, but as exemplified utilising the framework, it is possible to describe and analyse potential candidates for rational (partial) preference relations with the desired properties.

Location- and time-dependency is of specific interest in today's mobile society, where browsing long lists of answers is not feasible and more and more smartphone applications make use of GPS information to provide better user support. However, our studies indicate, potentially surprisingly, that the type of queries stated by mobile users of the health information portals may be less context dependent, that is, more related to information on diseases and other non-managerial aspects of the health care. Thereby the first principle would possibly be less of a problem for mobile health information seekers.

We also elaborated on the possibility to predict location-dependency based on a semantic analysis of queries, where ones corresponding to health care units, professions and certain types of activities may indicate specific interest in answers considering location information as part of the context used for a portal's interpretation and decision on answers to present. We estimate, based on automatic annotation of queries using the UMLS, that 20% of the interactions are location-dependent with 75% being explicitly chosen health care facilities, thereby further supporting the importance of Principle 1 for trustworthy portals.

In addition to location- and time-dependency, our analysis revealed potential problems with portal suggestions leading to seeker problems to find adequate answers, portal instructions which are often misunderstood by seekers, but also a substantial number of problems which seem to be of technical nature, and traces of a fluctuating interpretation behaviour in the sense that sometimes a query is interpretable and at another time non-interpretable within the same context, and this is not a feasible feature as it may lead to seeker distrust in the portal's way of acting. Hence, portals should as much as possible avoid bounded rationality with a short time span, since this may lead to seeker distrust.

Principle 5: A portal should avoid preference relations with bounded rationality and a short time span.

Another important aspect of the interactions between seekers and portals, and the former's trust in the portals, is how misspelt queries are managed. This is especially important since a seeker may be unaware of her posting to be incorrect. This is supported by our findings that, for instance, the disease borrelia is misspelt in 12.7% of all searches for Sweden. Moreover, today internet users are familiar with, and expecting, a search engine like Google which is able to manage queries that might be incorrect. Our analysis also noticed a tendency that the problem with misspelling might be more pronounced in regions and counties with a larger proportion non-native Swedish speakers, and many of the found deviations are similar to those being more common among people learning a new language. Hence, a fundamental problem with portals like 1177.se and vardguiden.se may be that the language they expect to be used is partly unfamiliar to the seekers, hence there is a risk for communication problems ultimately leading to distrust. In addition to Principle 2 this is captured by the following guideline:

Principle 6: The portal vocabulary should match the actually used seeker vocabulary, including acronyms, laymen terminology and possibly even medical terms known to be difficult to express correctly by the seekers.

A third interesting aspect of the studied portals is the way they treat semantically related information. For instance, that we found synonyms to be treated in different ways, where one term may result in answers and its synonym ends up with an empty answer set. Similarly, we found queries which resulted in answers at the same time as their hypernyms were considered non-interpretable. This is also a type of portal behaviour which could easily result in seeker distrust, viewing the portal as irrational an unpredictable, consequently, being captured by Principles 3 and 4 of a trusted portal.

The last aspect we studied was the occurrence of unknown queries, that

is, ones never being interpretable by the portal. Clearly this is a substantial problem without an obvious solution. However, as will be further discussed in the next chapter, there may be a relation between the queries which are difficult to annotate with semantic information to correspond to ones often resulting in specific types of answers to be of interest. If so, this may be used as a means to alleviate these problems. Moreover, we noted that mobile queries seem to be of potentially less risk to run into these problems, since these seekers might be more interested in information on diseases, which causes less problems for sources as the UMLS.

10 QUERIES WITH MANY ANSWERS

In chapter 9 we studied searches without answers, i.e. incomplete rounds in a search game. We also covered aspects related to interactions taking place using mobile devices and, as was noted in chapter 2, the importance of these will probably increase. Independent of if the use of health information portals take place by mobile devices or not, the seeker and portal share the interest in being able to satisfy their needs and responsibilities as well as possible. One such aspect is how to treat queries with too many possible answers to browse for mobile users. For instance, posting the query *feber hosta* 'fever cough' may indicate an interest in treatments as well as preventions, ending up with long lists of feasible answers for the portal to decide how to present. In the case of mobile searches, the seeker probably wants to receive just a handful of relevant suggestions.

In this chapter we address challenges related to the notion of "too many answers", based on the search log of Västra Götaland, in some cases in comparison to the county data in the Sweden dataset and mobile data from Stockholm. For Västra Götaland we have access to the posted queries, the number of provided answers, the rank of the chosen one and an indicator in the form of a hyperlink of the click. Hence, this dataset offers a possibility to study relations between queries and answers, and we are especially interested in if and how a portal provider could utilise such information to decrease the rank of certain types of query–answer pairs.

In section 10.1, we elaborate on the concept of "many answers" in the eyes of an information seeker, and based on the notion of *rank* of clicks we identify a set of queries possibly viewed as problematic for both seekers and portals, by often needing more than five answer gists to be considered before a seeker decides on the one to further pursue. In section 10.2, focusing on these so called *6-queries*, we present an overview of their *stylistics*, i.e. how the collective of seekers viewed as "authors" of a certain type of texts express themselves in terms of used terminology and linguistic constructs. This is followed by a presentation in section 10.3 along similar lines of the chosen answers with a focus

on a categorisation of the clicks to facilitate an analysis of relations among queries and answers in section 10.4. Hence, we show how important lessons may be drawn about the use of health information portals, and elaborate in section 10.5 on how they may lead to both a better understanding of the use of health information portals and a number of principles to establish and maintain seekers' trust in the portals.

10.1 6-queries – queries of mobile interest

When an information seeker posts a query, the result may be that she feels to be presented with a continuum from none to too many answers. In chapter 9, we studied the first extreme, where we concluded that many queries, more often than intended or wanted, led to incomplete search rounds without answers. In this chapter we turn our interest to the other extreme with queries resulting in large, and possibly diverse, answer lists. These two perspectives allow us, in chapter 11, to discuss what to be learnt from these types of interactions towards the aim of principles which should be adhered to by health information portals.

Even though this chapter is about queries with many answers, from an information seeker perspective, the number of presented answers is normally not a problem as long as the most interesting one is quite easy to find. The problem then turns into how to define "quite easy to find". A marker of this is the *rank* of the click, reflecting how many answer gists the seeker had to, in theory, read before deciding on one to be considered to, potentially, provide the information of interest. Hence, the higher rank the greater the risk of the portal not satisfying the needs of the seeker as easily as expected, and in the worst case being considered irrational and not to be trusted. Another marker could be to study the length of search sessions, and their content, but the available search logs lack in general accurate session information, and the overwhelming proportion of all sessions seem to consist of only one search round, see chapter 8.

In a search game, the rounds of interest are the ones where the captured answer situations, in terms of the rank of the queries, indicate a possibly poor portal preference relation for the given (type of) queries with their associated contexts. The first task is to identify these queries of interest, and to characterise the moves of seekers and portal leading up to these situations.

Figure 37 presents the distribution of the seeker vocabulary per average rank for Västra Götaland, and 69% of the queries resulted in an answer list where the seeker could find an interesting answer among the first three in the list. In other words, the average seeker will probably in one third of her moves end up with the portal making a move leading to the seeker having to read

more than three answer gists before finding an interesting one, or to spend three times more time to find an answer of interest in comparison to the, theoretically, ultimate case of Utopia with one answer per utterance. Moreover, 14% of the vocabulary required six or more answers to be browsed before an interesting one was found. Obviously, it may be that the seeker first checked the second and then the seventh answer, but due to unclear use of session identifiers this is difficult to judge. Another notable aspect is that, according to the log, there was never a need of more than 39 answers to possibly satisfy the seeker, and all except one were within 18 answers. The reason for this is probably that the rank only runs between the first and the last answer (e.g. 20) per page of answers (see section 8.1.1.6). Assuming that in the future the use of smartphones as search devices will increase, it is an interesting question if, and how much, it is possible to improve the portals towards the one query and one answer Utopia.

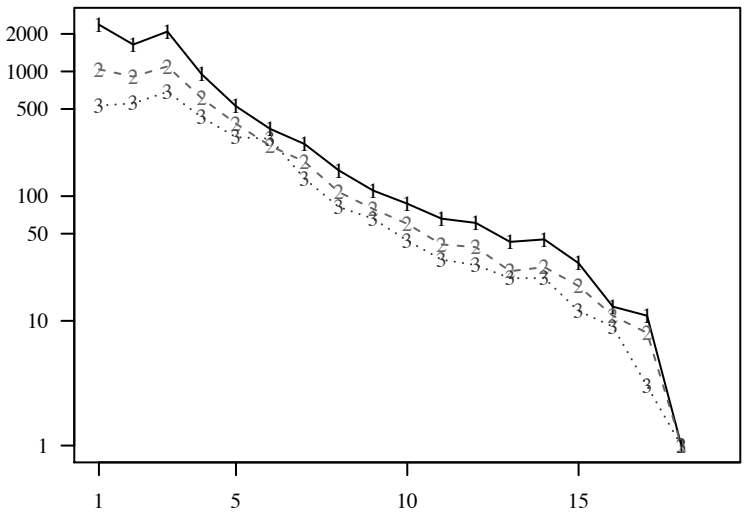


Figure 37: Distribution of seeker vocabulary per average rank for Västra Götaland (1), in comparison to those also found in the Stockholm Non-mobile (2) and Mobile (3) datasets.

Since for Västra Götaland we do not have information on if a search was mobile, figure 37 presents the number of queries with a given average rank for Västra Götaland also found for Stockholm Mobile and Non-mobile datasets, respectively. For instance, among the 526 queries with average rank five for Västra Götaland, 297 were also used for mobile searches in Stockholm, and 381 were also used in non-mobile searches. It is also worth mentioning that

less than 1% of the 3,147 potentially mobile queries were not used for non-mobile searches, hence according to this view, mobile searches do not differ with respect to the used queries in comparison to non-mobile. However, considering Stockholm Mobile versus Stockholm Non-mobile with the limitations above, this difference is considerably greater, but difficult to estimate due to a substantial amount of searches of unclear nature captured as complex technical records. By table 10.1 we also have that only very few of the more common mobile Stockholm queries would in the light of the Västra Götaland dataset lead to any problems to easily find answers of interest.

Query	**Average rank**
vårdcentral	2.8
barnmorskemottagning	3.4
feber	3.7
gynekolog	3.1
gravid	5.5
abortmottagning	2.0
förlossning	4.6
influensa	4.5
gynekologimottagning	–
halsfluss	3.9
akutmottagning	5.9

Table 10.1: Common (partially) interpretable mobile queries for Stockholm, and average rank of the corresponding queries for Västra Götaland.

Based on this we hypothesise that half of the Västra Götaland 6-queries were potential mobile queries at the time of posting, in comparison to a third of the ones with a lower rank. Thereby, 6-queries is of a potential interest from a mobile perspective.

Before continuing, it is important to comment on the use of data from different portals, time periods and regions to make statements about one of the datasets. Obviously, all the differences between the Västra Götaland and Stockholm datasets make any comparison questionable, but due to the lack of mobile information for Västra Götaland this is the only way to try to identify potential Västra Götaland mobile queries, and how difficult, measured in terms of the rank of the clicked answers, it might have been for a mobile seeker to find the answer of interest. Moreover, by our analysis mobile seekers often tend to end up with having to browse more than five answers, and obviously this can be questioned based on the origins of the data etc. However, it is important to remember that our aim of the use of both datasets is to try to identify some

examples of interactions of potential interest both from a mobile and usability perspective, and to characterise these as a basis for further analysis.

Based on the discussion above, we will study queries with an average rank greater than five, thereby implicitly asking the question what is needed to let a portal present only five answers to any query to improve portal usability, and ultimately the trusting beliefs in the portal by considering it competent and predictable.

10.2 Query stylistics

In chapter 9, we used the work by James (1998) on second language learning as a basis to study queries without answers, implicitly assuming many of the considered queries to result from information seekers not knowing how to express themselves in the language of "health care". In this chapter, we focus on the cases where the seeker was able to obtain answers, but possibly too many by expressing herself too vaguely, or the portal facing problems to find out the seeker's interests. Hence, we could say that the challenge might not only be for the portal to make use of the query and context as such in the best possible way, but also to consider aspects related to the way the seeker expressed herself. For instance, if she makes use of general medical terms like *cancer* or specific ones like *cervical cancer*, indicating a more focused interest and possibly also some knowledge on the topic. Another example is if the seeker uses terms, e.g. *vaccination*, which may indicate an interest in preventive actions or ones like *feber hosta* 'fever cough' implicating an interest in treatments. It is not only important for a health information portal to consider *what* is typed by a seeker, but also *how* it is done, that is the "style" of the queries.

When we studied different types of problems leading to information seekers not obtaining any answers, following James (1998), we described deviations at the levels of substance, text and discourse. Focusing on the information seeker's query, the substance level would then correspond to the actual *queries* of the seeker, the text level to the grammatical and lexical *rules* and *patterns* and the discourse reflects the *intentions* of the seeker. As discussed in the theoretical introduction to this section (section 6.6), from a stylistics perspective the division of James also makes sense in settings where focus has shifted from deviations to interpretable utterances.

With the categorisation of query aspects into the substance, text and discourse points of view, we can at a substance level, in addition to measures as *query length*, assuming longer queries to lead to fewer answers and lower ranks, also study stylistic features such as the use of *compounds*, providing insights into their role of constraining interest to certain types of care or health

care units, *capitalisations* and *acronyms*, denoting diseases and types of care givers, and *conjunctions* to define complex queries. Aspects like the type of language and use of lexicological relations like *synonymy* and *hypernymy* of the semantics of the terms is then studied at text level, allowing us to address questions like which type of queries that are more common and how detailed seekers express their interests. These aspects are interesting since we may assume more detailed queries to lead to fewer answers and lower rank, see for instance (Belkin et al. 2003), and synonyms to result in no differences with respect to these aspects. It is also interesting to study if certain types of queries, e.g. ones related to health care management, more often result in higher ranks than, for instance, ones related to diseases. Since our focus is on queries of interest to mobile search, especially so called 6-queries, the studies will share the theme on if and how these aspects impact portal behaviour. In section 10.4 we study these in the light of the discourse of the interactions as given by the query–answer relations, thereby being able to study patterns of interactions reflecting different types of search scenarios and if some are more likely to result in potential seeker dissatisfaction.

The choice of features to study is a subjective decision, and mainly based on our own findings when exploring the search logs. However, we are also inspired by questions addressed in literary stylistics where the aim is to identify variations and common patterns among (famous) authors, but in our case we view the collective of health information seekers as the "author" of interest.

To summarise, the aim of this section is to describe some *characteristics* of queries often leading to more than five answers for an information seeker to browse before considering one interesting, and with special attention to the queries which might be more prevalent among mobile searches, and how these may help information providers decide on answers to present to reduce the effort of the information seekers to find adequate support. Hence, the focus is on the seeker's move and its characteristics, keeping in mind that the query and the context are the only available features to the portal in its mission to provide the best move and establish and maintain the seeker's trust, solely based on a rational preference relation among answers.

10.2.1 Substance

When we studied substance deviations, we saw that they were made up of those related to misspellings resulting from "mechanical" problems such as punctuation and typographic ones and proper misspellings possibly originating in lack of knowledge of the used language, e.g. Swedish spelling of *colon* as *kolon* or the use of grapheme "ng" wrongly thought to represent the "target" sound "g"

in *angora fobi* (intended expression was probably *agorafobi* 'agoraphobia').

When considering queries with obtained answers, the substance level will not deal with misspellings, but with spelling patterns and variations as the use of capitalisation to indicate places and compounds to refer to health care units and professions. Hence, we will study the existence and use of *substance stylemes* of health information interactions as found in the use of health portals like 1177.se and vardguiden.se.

10.2.1.1 Queries

Tables 10.2 and 10.3 present the most common queries for Västra Götaland and interpretable ones for Västra Götaland County.[64] Even though semantic aspects of queries will be discussed in section 10.2.2, it is worth noticing the difference in queries between the two datasets, with queries related to health care management, such as *mina vårdkontakter* 'my care contacts' and *vårdgaranti* 'health care guarantee', being more common for Västra Götaland. There may be different reasons for this, but one could be the newer interface used for the county to better manage this type of concerns, thereby less need for explicit queries. Another reason could be that today's information seeker is better informed on these matters. Hence, the differences in queries as such indicate a potential difference in the user *interfaces* of the portals and how certain types of *needs* are cared for by the providers.

Query	**Concept**	**Semantic type**
mina vårdkontakter	Patient Care Management*	Health Care Activity*
egenremiss	Self-referral*	Health Care Activity*
sjukresor	Benefits, entitlements and rights*	Intellectual Product*
vaccination	Vaccination	Therapeutic or Preventive Procedure
gravid	Patient currently pregnant	Finding
urinvägsinfektion	Urinary tract infection*	Disease or Syndrome*
feber	Fever	Finding
influensa	Influenza	Disease or Syndrome
vårdgaranti	Benefits, entitlements and rights*	Intellectual Product*
byta vårdcentral	Patient Care Management*	Health Care Activity*

Table 10.2: Common queries for Västra Götaland.

[64] In this chapter, concepts and semantic types marked with an asterisk (*) denote ones manually assigned based on our interpretation of the query. See also discussion on automatic annotation in section 8.2.3.

Query	Concept	Semantic type
gallsten symptom		Sign or Symptom*
tjocktarmscancer	Malignant tumor of colon*	Neoplastic Process*
svullna fötter och ben	Swollen legs*	Sign or Symptom*
stroke symtom		Sign or Symptom*
husläkare	General practitioners*	Professional or Occupational Group*
dalslands sjukhus	Health Care Facility*	Health Care Related Organization*
stress symtom	Symptoms of stress*	Sign or Symptom*
hjärnan	Brain*	Body Part, Organ or Organ Component*
tandläkare	Dentist	Professional or Occupational Group
röntgen	Plain x-ray	Diagnostic Procedure

Table 10.3: Interpretable queries for Västra Götaland County.

If we consider the 6-queries for Västra Götaland, table 10.4, we have that in almost half of the cases the most common terms are about management and, together with tables 10.2 and 10.5, these queries are also among the more common ones in general, thereby highlighting the risk of seekers to have to spend substantial time to identify answers of interest.[65] In addition, among the 6-queries we find ones like *hosta* 'cough', *magkatarr* 'gastritis' and the tick-related diseases *borrelia* and *tbe*.

It is difficult to interpret these results, but the first two refer to rather common problems and the latter ones to a topic of general public interest often discussed in media. Commonality and public interest are features shared also by *feber* 'fever' and *influensa* 'influenza'. Hence, in addition to information on if, and how well, certain interests are cared for by the portal interface, the commonality and public interest of 6-queries may also indicate general knowledge seeking to result in higher ranks, possibly by seekers "browsing" without clear aims using terms of a more "vague" nature, e.g. symptoms. Table 10.5 also shows that it seems like queries related to specific diseases like *diabetes* and *urinvägsinfektion* 'urinary tract infection' more seldom lead to problems to easily find satisfying answers, than queries related to specific administrative topics. As will be presented in sections 10.3 and 10.4, the portal structure is roughly divided into pages on diseases, management and different themes.

[65] In this chapter, queries marked with a dagger (†) denote ones not found among the Stockholm Mobile queries.

Query	Concept	Semantic type
mina vårdkontakter	Patient Care Management*	Health Care Activity*
egenremiss	Self-referral*	Health Care Activity*
vaccination	Vaccination	Therapeutic or Preventive Procedure
vårdgaranti	Benefits, entitlements and rights*	Intellectual Product*
mammografi	Mammography	Diagnostic Procedure
hosta	Coughing	Sign or symptom
tbe	Tick-Borne Encephalitis	Disease or Syndrome
magkatarr	Gastritis	Disease or Syndrome
mvc	Antenatal clinic*	Health Care Related Organization*
borrelia	Borrelia	Bacterium

Table 10.4: Common 6-queries for Västra Götaland.

Query	Rank >5	Query	Rank >5
mammografi	58.5%	sjukresor	36.6%
egenremiss	56.6%	diabetes	35.4%
mina vårdkontakter	52.3%	halsfluss	22.3%
hosta	50.8%	vattkoppor	17.7%
klamydia	47.7%	stroke	17.0%
vaccination	45.5%	feber	15.7%
influensa	43.9%	urinvägsinfektion	14.1%
vårdval	43.7%	blanketter†	10.9%
vårdgaranti	42.9%	webbisar†	0.3%
gravid	41.2%	byta vårdcentral	0%

Table 10.5: Common queries and proportion of search rounds with clicks of rank higher than five for Västra Götaland.

Moreover, the parts on diseases are generally on the form of fact sheets with symptoms, treatment etc, but a similar organisation is not found for health care management topics. Hence, queries related to specific diseases may thereby end up with lower rank than others. However, it is interesting to note the difference in choice of answers between two rather vague symptoms such as *feber* 'fever' (15.7%) and *hosta* 'cough' (50.8%), where the latter more than three times as often lead to longer time spent to find adequate answers. As will be seen in section 10.4, one explanation for the difference in challenges finding adequate answers for cough in comparison to fever may be that in the latter

case there is a thematic portal page on the topic.

From a search game perspective, this indicates that when the seeker's first move, i.e. the utterance, is related to aspects like *browsing* on topics of public interest, *uncertainty* by posting terms referring to general symptoms, and *usability* problems related to, for instance, administrative matters not covered clearly by the portal, it is more likely to result in portal preference problems as seen by higher rank of the chosen answers. This aspect is important, since by the role of the portal as an official public mediator of information on health and care, a seeker probably approaches it with a disposition of trust and trusting intentions. Thereby, usability problems and browsing not leading to increased certainty may decrease the beliefs in the portal and ultimately a potential distrust. Hence, in these cases further attention by the portal to the context or historical data on query–answer relations may provide insights allowing the portal to better address the seeker needs, especially regarding patterns of browsing on topics of public interest and interactions on ones considered to be covered by static parts of the portal, such as dedicated areas of care management.

10.2.1.2 Query length

A reasonable assumption regarding the portal preference relations is that the more query terms, including context, provided by the seeker the fewer the answers which would be of potential interest to the seeker. In more sophisticated solutions, it is fair to assume that longer queries constrain the answer list resulting in the more focused answers to end up higher in the list of presented ones.

Figure 38 presents the distribution of lengths of the seeker vocabulary for Västra Götaland, showing that there is no difference between queries in general and 6-queries, thereby not confirming a hypothesis of longer queries to be more constraining and lead to more detailed answers and easier to find interesting ones among the presented alternatives. The figure also shows that in the case of queries found in Stockholm Mobile there might be a weak tendency of mobile queries to be shorter in comparison to queries in general and 6-queries. Studying the number of answers and query length for Västra Götaland and the county, figures 39 and 40, there might be a minor trend, especially for the county, of longer queries to lead to in average fewer answers. However, the differences are, in our opinion, rather small in comparison to what might be expected when the length of queries is doubled. The figures also show that in general the average size of an answer list is in the range of 10–50, but there are many with over 50 answers.[66] If we also consider the rank, figure 41, going

[66]It is difficult to explain how queries end up with thousands of answers, since they do not

from one to two query terms may slightly reduce rank, but in the rest of the cases query length does not seem to have an impact on the process of finding an answer of interest.

To summarise, many seekers might expect more detailed queries, as measured by query length, to lead to fewer answers, but above all, fewer answers to have to consider before finding one of interest (Belkin et al. 2003). However, this does not seem to be the case and this may lead to seeker distrust with respect to how the portal's preference relation among answers functions and how seekers should phrase their queries to affect its outcome. Hence, it impacts the seekers' beliefs in the portal, and ultimately a trust-related behaviour. Moreover, if the portal tries to adapt to a change in seeker behaviour, this may even lead to further seeker problems and distrust.

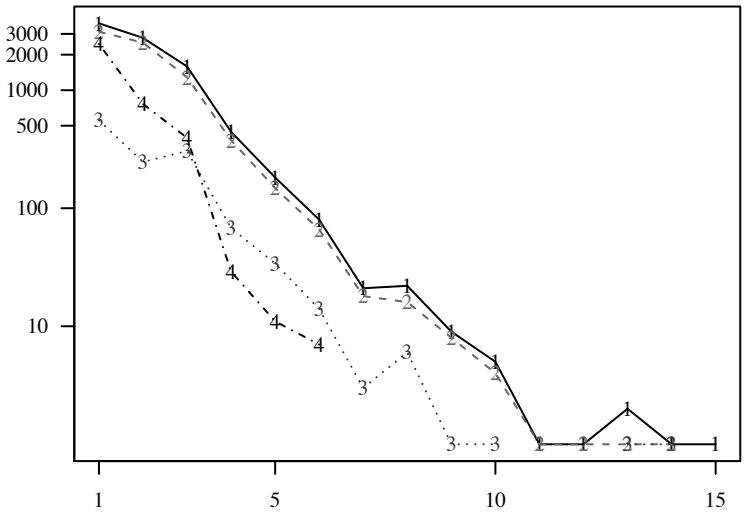


Figure 38: Distribution of lengths of queries (1), non-6-queries (2), 6-queries (3) and potentially mobile queries (4), respectively, for Västra Götaland.

show any deviating patterns in comparison to other queries.

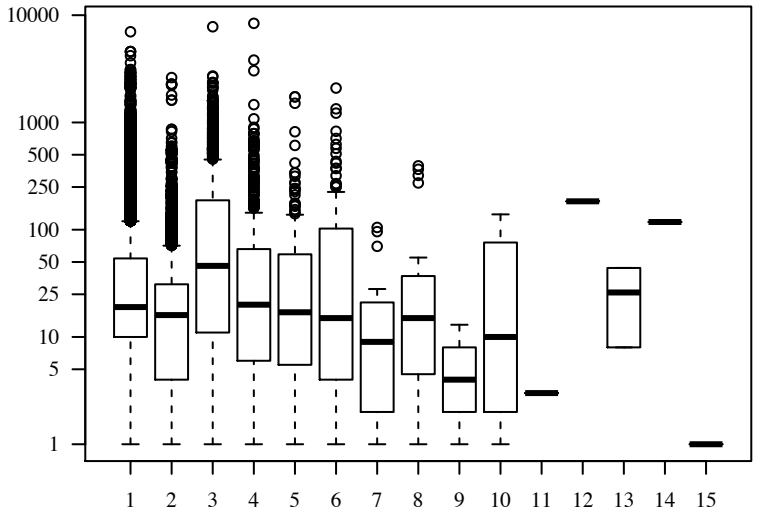


Figure 39: Number of answers (y) versus query length (x) for Västra Götaland.

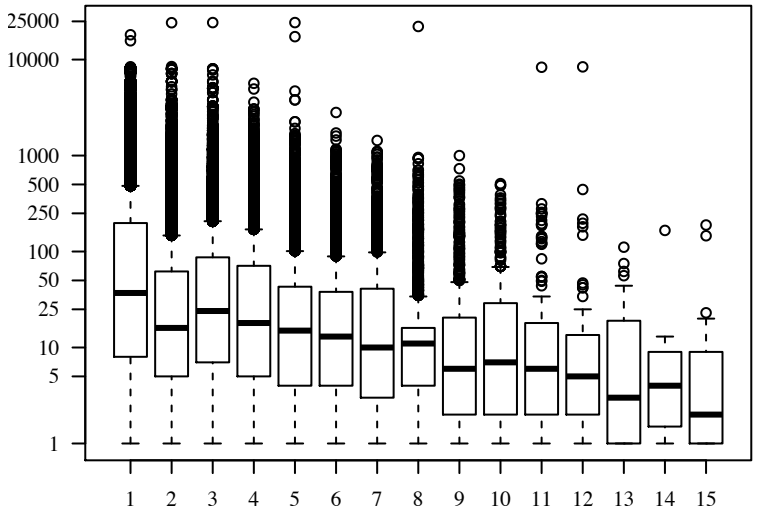


Figure 40: Number of answers (y) versus query length (x) for Västra Götaland County.

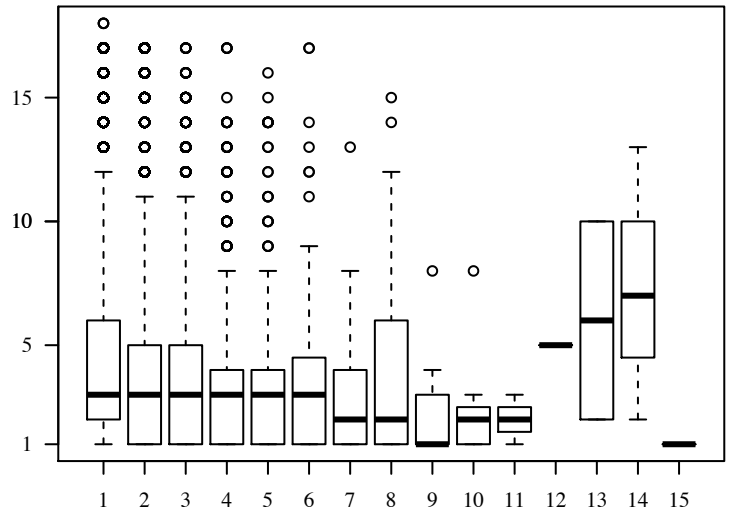


Figure 41: Rank (y) versus query length (x) for Västra Götaland.

10.2.1.3 Compounds

Already in chapter 9 we saw that information seekers quite often tend to use compound terms, e.g. *husläkare* 'general practitioner'. Hence, an understanding of their use, especially in the light of 6-queries, may be valuable. The most interesting of these may be the ones called *endocentric* with more than one stem where one is the head and the rest restrict, or modify, its meaning (section 6.6.2). For instance, *tandvård* 'dental care' where *vård* 'care' is the head and *tand* 'dental' restricts the interest to a certain type of care, consequently potentially limiting the seeker interests or indicating an interest in the topic of the head stem of the compound. This potential impact of compounds as limitors of answers, and more importantly facilitators of lower ranks, is interesting, and table 10.6 presents some common compounds. In this case almost half of the ones for Västra Götaland deal with health management, but except for *egenremiss* 'self-referral' and *vårdgaranti* 'health care guarantee' they do not seem to cause any problems for the seekers in finding answers among the first ones presented by the portal.

Västra Götaland		**Västra Götaland County**
Query	6-query	Query
egenremiss	egenremiss	vårdcentral
sjukresor	vårdgaranti	jourcentral
urinvägsinfektion	magkatarr	tandvård
vårdgaranti	bröstcancer	akutmottagning
magsjuka	svininfluensa	sjukhus
lunginflammation	sjukgymnastik	barnavårdscentral
tandvård	glutenintolerans	barnmorskemottagning
högkostnadsskydd	bensår	magsjuka
njursten	förmaksflimmer	lunginflammation
sjukvårdsupplysningen	halsböld	preventivmedel

Table 10.6: Examples of compound queries for Västra Götaland and Västra Götaland County.

It is also interesting that rather well-defined terms such as *glutenintolerans* 'gluten sensitivity' and *förmaksflimmer* 'atrial fibrillation' potentially demand more effort from the seeker. It is worth noticing that considering Västra Götaland County we find several queries related to *mottagning* 'surgery' and *central* 'centre', possibly indicating an increased interest in different types of health care units. Table 10.7 shows, as in table 10.5, that among common queries for Västra Götaland ones related to health care management more often tend to result in higher rank than ones related to diseases.

Query	Rank >5	Query	Rank >5
egenremiss	56.6%	gallsten	27.2%
magkatarr	46.0%	öroninflammation	26.3%
högkostnadsskydd	45.3%	sjukresa	20.9%
vårdval	43.7%	urinvägsinfektion	14.4%
vårdgaranti	42.9%	njursten	13.4%
tandvård	37.9%	hjärnskakning	13.0%
sjukresor	36.6%	väntetider	11.6%
patientavgifter	28.8%	magsjuka	10.5%
färdtjänst	28.6%	lunginflammation	1.4%
svinkoppor	27.9%	sjukvårdsupplysningen	0.0%

Table 10.7: Common compounds with proportion of search rounds with clicks of rank higher than five for Västra Götaland.

Compounds, especially the endocentric ones, may be expected to result in fewer answers and lower ranks than other query terms. However, since many of the terms in the seeker and portal vocabularies tend to be compounds it is difficult to judge the impact of this substance feature. The occurrence of these being 6-queries related to health care management may be a result of general portal challenges with this type of seeker interests rather than due to being compound queries, cf the discussion in the introduction of this section on common queries.

A related question addressed in section 10.2.2, is how the portal preference relation for compounds relate to its treatment of hypernyms, for instance the compound *bröstcancer* 'breast cancer' in relation to the hypernym query *cancer* where we might expect hypernyms to more often lead to higher ranks than hyponym endocentric compounds. An interesting example for Västra Götaland County is the query *adenokarcinom* 'adenocarcinoma', which stands for cancerous (carcin-) tumour (-oma) of glands (adeno-). This term results in no answers, but the hypernym cancer is interpretable. This example shows the importance of a portal to be able to manage compounds in the light of their onomasiological environment.

10.2.1.4 Capitalisation

In Swedish the use of capital letters, *capitalisation*, often denotes proper nouns and abbreviations formed by the initial letters of the components of a word or phrase, called acronyms. Table 10.8 shows examples of capitalised query terms for Västra Götaland. The capital letter is either the first letter in a query or part

of an acronym. In the first case it may result from the use of smartphones or tablets often configured to always use initial capital letters. Hence, identification of more than one capital letter in a query may be used to recognise (medical) acronyms used by information seekers.

Query	**6-query**
Mina vårdkontakter	Vaccination
TBE	ADHD
Urinvägsinfektion	Mammografi
Sjukresor	Gravid
KOL	Njurbäckeninflammation
MS	MVC
BUP	NÄL
Vaccination	Nässelutslag
Vårdgaranti	LERUM
ADHD	Mariestad

Table 10.8: Examples of capitalised queries for Västra Götaland.

For Västra Götaland we have that 6.2% of the queries (6.4% for ones not being 6-queries) begin with a capital letter which is followed by a lowercase, and 1.5% of the queries of length one have more than one uppercase letter (1.4% only capitals). In the case of 6-queries, these numbers are 4.8% and 1.6% (1.5% only capitals), respectively, indicating a potentially increased use of acronyms leading to problems finding adequate answers. If we study queries with more than one term, in 5.1% of the cases (5.5% for ones not being 6-queries) the second word begins with a capital letter, 1% with only capitals, and 4.0% (4.5% for non 6-queries) have a second word beginning with a capital letter followed by a lowercase letter. This is to be compared to 6-queries where the numbers are 2.6% (1.5% only capitals) and 1.0%, respectively. In other words, queries not being 6-queries with at least two terms are ten times more likely than 6-queries to have a second term with at least two capital letters, implicating capitalisation in the second term to play a role in the resulting rank. If we study these queries, we find a higher degree of diversity among the topics covered by the 6-queries, often generally difficult for the portal to manage, but in the case of queries not belonging to the category of 6-queries we find ones related to specific health care units. It is also worth noticing that 1.2% of all queries, and 1.4% of the 6-queries, consist of only capital letters supporting the thesis of acronyms to more often lead to higher ranks.

To summarise, from a portal perspective the seeker moves can be divided into those where capitalisation may indicate acronyms or proper nouns like

names of care givers or locations, and those where it is probably of less significance. Based on our analysis, it seems that the acronyms have a tendency to more often lead to problems in easily finding interesting answers, and this topic is discussed in the next section.

10.2.1.5 Abbreviations and acronyms

As exemplified in the previous section, in the language of medicine, *acronyms*, i.e. abbreviations formed by the initial letters of the components of a word or phrase, are often used to denote laboratory procedures, diseases or health care units. Table 10.9 presents common abbreviations including acronyms for Västra Götaland, for instance *tbe* (tick-borne encephalitis) and *ms* (multiple sclerosis*)*.

Västra Götaland		**Västra Götaland County**
Query	6-query	Query
tbe	tbe	ms
ibs	mvc	mvc
mvc	bup	ibs
hiv	sle	hiv
adhd	bb	adhd
ms	bvc	tbe
als	kss	als
bup	hpv	bvc
sle	su	bup

Table 10.9: Examples of abbreviations and acronyms for Västra Götaland and Västra Götaland County.

Table 10.10 shows indications of a tendency of abbreviations related to health care units such as *mvc* 'antenatal clinic' and *bup* 'child and adolescent psychiatry' to more often lead to ranks higher than five, but it is also surprising to find well-defined concepts like *ibs* (irritable bowel syndrome) and *sle* (systemic lupus erythematosus) among the ones which rather often result in higher ranks.

However, if we compare the average ranks of *mvc* and its non-abbreviated form *mödravårdscentral* 'antenatal clinic' for Västra Götaland, there is no substantial difference. This also holds for *ms* and its non-abbreviated form, which both have an average rank between three and four. Hence, it is difficult to draw any conclusions on the use of abbreviations based on the available search logs, and the examples may be more indicative of a general portal problem with

Query	Rank >5	Query	Rank >5
bup	62.1%	adhd	23.3%
sle	51.8%	ivf	14.0%
mvc	51.6%	hiv	13.6%
bvc	50.8%	näl	13.0%
bb	43.8%	kbt	8.7%
tbe	39.1%	ms	3.5%
ibs	37.1%	als	0.0%
gyn	33.3%	npö†	0.0%
tia	31.0%	uvi	0.0%

Table 10.10: Common acronyms with proportion of search rounds with clicks of rank higher than five for Västra Götaland.

certain types of topics independent of potential abbreviations.

Acronyms are intended to provide short-form "names" of well-defined concepts, hence, in theory, be viewed as high-qualitative expressions for a portal when deciding its moves. Moreover, these queries should be considered as good as their non-abbreviated form. Since abbreviations and acronyms, especially as denotations of rather complex phrases, are common in health related terminology, and sometimes the only term known by a seeker to denote a concept, a portal would probably benefit from ensuring ability to interpret the short forms as well as the non-abbreviated forms.

10.2.1.6 Conjunctions

In information search, words like *och* 'and' and *eller* 'or' are often used as conjunctions in combined queries, and table 10.11 shows some examples.[67] By the complexity of the queries, we believe they may be the result of the seeker choosing a query suggested by the portal, which in almost all cases results in the seeker having to consider in average at least six different answers before finding one of interest. Independent of the use of conjunctions, from a mobility perspective, this highlights the possibility of a portal to steer the seeker towards certain answer lists, but also that this may lead to seekers having to spend more efforts on finding the interesting information. Moreover, the conjunction *eller* 'or' was used in only 14 cases, for instance *av- eller omboka tid* 'cancel or reschedule appointment'.

[67]The use of prepositions and conjunctions in health information search has also been studied by us in (Eklund 2012b).

Västra Götaland	
Query	6-query
mag och tarm	mag och tarm
om eller avboka tid till hälsounders	barn- och ungdomshabiliteringen
preventivmetoder och preventivmedel	hjärt och kärlsjukdomar
snuva och hosta	kräkningar och illamående
alkohol och träning	ont i händer och fötter
barn- och ungdomshabiliteringen	smärta och smärtlindring
domningar och yrsel	magont, huvudvärk och feber
drottning silvias barn- och ungdomssjukhus	ont i bröstet och hosta
kräkningar och illamående	synskada och samhälle
ont i händer och fötter	ont i vänster axel och mage

Västra Götaland County
Query
barn- och ungdomsmedicin
t.ex. mottagning eller typ av vård
sex och samlevnad
borrelia symptom och behandling
yrsel och illamående
trötthet och orkeslöshet
svullna fötter och ben
yrsel och huvudvärk trötthet och illamående
feber och hosta
illamående och trötthet ej kräkningar

Table 10.11: Uses of conjunctions for Västra Götaland and Västra Götaland County.

From a search game perspective, conjunctions indicate how a portal is to interpret the query, as a logical conjunction (and) or logical inclusive disjunction (or). In the first case, this normally corresponds to the implicit way portals interpret queries with several terms, i.e. as a sequence of terms each to be satisfied by the answer. However, logical disjunctions invite the portal to provide answers which satisfy at least one of the combined terms. Hence, this type of construction adds further possibilities for seekers to indicate uncertainty, but by our analysis this feature is seldom used today.

As a side note, the finding of potential portal suggestions to lead to seeker challenges to find interesting answers indicate a fundamental problem with maintaining a seeker's trust who willingly depends on the portal. Hence, if the portal is to try to "steer" a seeker (cf the example and discussion in section

5.2.1), it has to make sure not to jeopardise the seeker's trusting intentions.

10.2.1.7 Summary

From a search game perspective, our substance analysis indicates that when the seeker's move is related to aspects like *browsing* on topics of public interest, *uncertainty* by posting terms referring to general symptoms, and *usability* problems related to, for instance, administrative matters not managed by other parts of a portal, it is more likely to result in portal preference problems as seen by higher rank of the chosen answers. Moreover, neither longer queries, compounds nor acronyms seem to lead to improved abilities for seekers to easily find adequate answers. Since the relation between health information seekers and portals needs to be based on trust, only materialised by queries and answers, many of our findings indicate potential problems and possible improvements even without involving sophisticated text level analysis based on semantic interpretation of queries and answers.

10.2.2 Text

Text level analysis considers aspects mainly related to the semantics, or meaning, of the queries, for instance, queries like *feber* 'fever' to be a *Finding* indicating that the seeker is possibly suffering from a disease she wants to treat. We focus on two perspectives addressing the questions of how seekers express certain health care *interests* and which *needs* are expressed by the queries. These are called *onomasiological*, i.e. how concepts are expressed by terms, and *semasiological*, i.e. which concepts are expressed by given terms.

Seekers use different terms to express their needs and sometimes many of them refer to the same concept, hence they are *synonyms*. Understanding the use of synonyms allows us to study if a portal is able to manage different ways to express a concept as expected by seekers, that is, provide the same answer list for the use of any synonym, thereby a consistent and rational portal behaviour in the eyes of the seeker. Other interesting aspects are if there is a common use of *polysemy*, i.e. one term which expresses more than one concept, thereby leading to portal challenges to decide adequate answers, and at how many levels of detail concepts are used to express the needs, that is, the degree of *hyponymy* in the searches and if it differs with respect to type of concept. In this case, in the eyes of the seeker, the more detailed query should result in fewer answers and lower rank of the ones of interest. However, we will start our analysis with the question of which concepts seekers find interesting.

10.2.2.1 Semasiological relations

Since the Västra Götaland vocabulary consists of almost 9,000 queries, it is impossible to manually analyse the concepts expressed by these and we have to rely on annotations using the sources of the UMLS (sections 6.6.3 and 8.2.3). However, with a basic look-up of queries in the UMLS, only in the range of 20% of the Västra Götaland seeker vocabulary can be automatically assigned concepts.[68] It is also worth noticing that only around 6% of the (partially) interpretable queries for the county can be mapped to concepts.[69] If we constrain our interest to the 42% of the vocabulary for Västra Götaland which consist of only one term, the proportion of mappable ones is 46%. Hence, the ability to semantically annotate queries is a major challenge, see also the discussion in section 9.4 on unknown utterances. Tables 10.2 and 10.4 showed examples of common queries and their concepts, and if we categorise the mappable ones based on their semantic type, *Disease or Syndrome* is clearly the most common type of concepts expressed by these queries, tables 10.12 and 10.13. Moreover, considering the ratio of queries versus proportion of the mappable seeker vocabulary, there is a substantially greater diversity in the queries related to *Pharmacologic Substance* and *Organic Chemical*, e.g. drugs and vitamins, and *Body Part, Organ, or Organ Component* in comparison to ones related to *Disease or Syndrome*. If we focus on the concepts most often resulting in ranks higher than five, we find ones of the semantic types *Health Care Activity*, *Therapeutic or Preventive Procedure* and *Sign or Symptom* and in general slightly greater diversity in comparison to queries related to diseases.

Furthermore, by table 10.14 the mappable 6-queries for Västra Götaland in addition to the above, indicate that even though there is greater diversity in queries related to drugs, vitamins and body parts, they seem not to be difficult to find adequate answers for. At the same time, queries related to procedures and symptoms tend to be more difficult for the portal to manage. This may support the hypothesis that concepts with well-defined names and semantics are easier for the portal to manage than ones considering more vague ones like like symptoms and procedures.

[68] The use of semantic annotations of queries has also been studied in (Eklund 2012c) in the context of layman versus professional language.

[69] Problems with mapping terms to disorders and findings using Swedish Snomed CT in the context of clinical text have been studied by Skeppstedt, Kvist and Dalianis (2012).

Semantic type	**Mappable seeker vocabulary**	**Utterances**	**6-query utterance**
Disease or Syndrome	21.9%	4,534	11.0%
Pharmacologic Substance	15.5%	1,099	15.3%
Finding	5.6%	1,097	16.3%
Therapeutic or Preventive Procedure	7.1%	946	37.5%
Sign or Symptom	4.7%	898	32.4%
Organic Chemical	11.5%	751	16.1%
Body Part, Organ, or Organ Component	5.5%	507	19.1%
Intellectual Product	1.8%	429	4.4%
Manufactured Object	1.7%	376	1.6%
Health Care Activity	2.2%	367	39.8%
	Total: 77.5%	Total: 11,004	

Table 10.12: Most common semantic types of mappable queries for Västra Götaland.

Semantic type	**Mappable seeker vocabulary**	**Utterances**
Disease or Syndrome	51.8%	226,628
Health Care Related Organization	0.8%	102,887
Manufactured Object	2.4%	92,965
Pharmacologic Substance	45.2%	56,435
Organic Chemical	28.9%	40,195
Sign or Symptom	8.5%	38,879
Finding	9.2%	37,778
Neoplastic Process	7.7%	25,942
Body Part, Organ, or Organ Component	19.0%	21,572
Therapeutic or Preventive Procedure	9.7%	20,635
	Total: 183.2%	Total: 663,916

Table 10.13: Most common semantic types of mappable queries for Västra Götaland County.

Semantic type	**Mappable seeker vocabulary**	**Utterances**
Disease or Syndrome	11.5%	499
Therapeutic or Preventive Procedure	7.8%	355
Sign or Symptom	7.4%	291
Finding	10.5%	179
Diagnostic Procedure	3.4%	169
Pharmacologic Substance	8.5%	168
Pathologic Function	5.1%	155
Health Care Activity	4.1%	146
Organic Chemical	5.1%	121
Body Part, Organ, or Organ Component	8.5%	97
	Total: 71.9%	Total: 2,180

Table 10.14: Most common semantic types of mappable Västra Götaland 6-queries.

One explanation for this could be that the portal is organised around "hard" facts of diseases, drugs and anatomy to a greater extent than with respect to symptoms and procedures. However, as seen by tables 9.24 and 9.25 (section 9.4), many of the non-mappable queries are related to health care management, and this is an area not addressed as well as, for instance, diseases. These aspects are further discussed in section 9.4, and in the rest of this section our interest is focused on the types of queries found in the summaries in tables 10.12 and 10.13.

An interesting question is if it is common with queries with several different interpretations, i.e. different semantic types in the terminology of the UMLS, or polysemy. Table 10.15 shows examples of queries which may be associated with more than one semantic type, but the risk of resulting in higher ranks is small, by 5.0% of the vocabulary of 6-queries and 5.6% of queries in general for Västra Götaland having several types.[70] Hence, portal moves resulting from incorrect semantic interpretation of terms should be at most around 5% of the vocabulary, and considering the diversity not be a major problem.

To summarise, the analysis of the Västra Götaland search log reveals that in the case of unknown utterances (section 9.4) automatic semantic annotation of queries is a challenge, but based on the around 20% of the seeker vocabulary consisting of queries mappable to a UMLS concept, corresponding to 50.1% of the search rounds, we have that queries related to symptoms and procedures are especially difficult for the portal to manage. Hence, any semantic-based portal preference relation may be challenged with trust problems for topics related to,

[70] Queries marked with a Yen sign (¥) denote 6-queries.

for instance, symptoms which can be hypothesised to originate in seekers with uncertainty of what they suffer from and how to treat the cause. Moreover, the uncertainty is also related to the seekers "depending" on the portal to "solve" their problems, and inadequate support may severely damage the trust in the portal and even the seekers' health.

Query	**Concept**	**Semantic type**
depression	Depressed mood	Finding
	Depressive disorder	Mental or Behavioral Dysfunction
	Sad mood	Mental Process
insulin	Insulin	Amino Acid, Peptide, or Protein
	Insulin	Pharmacologic Substance
	Insulin	Hormone
sjukhus (¥)	Hospitals	Health Care Related Organization
	Hospitals	Manufactured Object
	Hospital environment	Qualitative Concept
blod	Blood	Tissue
	peripheral blood	Body Substance
penicillin	Penicillins	Organic Chemical
	Penicillins	Antibiotic
rehabilitering (¥)	Rehabilitation therapy	Therapeutic or Preventive Procedure
	Rehabilitation service	Health Care Activity
sår (¥)	Ulcer	Pathologic Function
	Injury wounds	Injury or Poisoning

Table 10.15: Examples of queries with several semantic types for Västra Götaland.

10.2.2.2 Onomasiological relations

Concepts may be expressed by several different terms, or *synonyms*, and the use of synonyms might be seen as an indication of how familiar seekers are with a concept, but also how often it is used in the media etc. By table 10.16, synonyms are mainly used for *Disease or Syndrome*. However, it also shows that in many cases, e.g. gastritis, there is a substantial difference in considering the proportion of 6-queries among synonyms. This is interesting, since the information seeker may expect utterances expressing the same concept to result in the same portal interpretation and move. Since only 20% of the Västra Götaland vocabulary is mappable to UMLS concepts, the 6.4% with synonyms may be a too low estimate of the importance of synonym treatment.

Another indication of how well a portal treats (semantically) related queries

Query	Utterances	Rank >5 (%)	Concept	Semantic type
halsfluss/tonsilit	148/3	22.3 / -	Tonsillitis	Disease or Syndrome
magkatarr/gastrit	69/12	46.0 / 16.4	Gastritis	Disease or Syndrome
nässelutslag/nässelfeber/urtikaria	36/10/1	34.1 / - / -	Urticaria	Disease or Syndrome
cysta/cystor	29/12	45.2 / 41.7	Cyst	Disease or Syndrome
hjärnhinneinflammation/encefalit	26/2	18.5 / -	Encephalitis	Disease or Syndrome
tuberkulos/tbc	5/15	- / 26.7	Tuberculosis	Disease or Syndrome
krupp/falsk krupp/pseudokrupp	4/15/1	-/ - / -	Croup	Disease or Syndrome
sår/ulcerös/ulcus	16/1/2	30.8 /- / -	Ulcer	Pathologic Function
hes/heshet	4/12	- / 25	Hoarseness	Sign or Symptom
crohns sjukdom/crohn	9/3	- / -	Crohn Disease	Disease or Syndrome
gastric/gastro/mage/magen	3/1/2/3	- / - / 50.0 / 25.0	Stomach	Body Part, Organ, or Organ Component
basalcellscancer/basaliom	4/3	- / -	Basal cell carcinoma	Neoplastic Process
bukspottkörtel/pancreas	5/1	44.4 / -	Pancreas	Body Part, Organ, or Organ Component
ljusbehandling/ljusterapi	3/3	- / -	Phototherapy	Therapeutic or Preventive Procedure
synnedsättning/nedsatt syn	2/1	- / -	Vision, Low	Disease or Syndrome

Table 10.16: Examples of synonyms for Västra Götaland

is to study the use of hypo- and hypernyms, and by table 10.17 we see that, as expected, it is more difficult for the portal to decide adequate answers in the case of inflammation in comparison to the more specific tonsillitis. However, it is also more difficult for the portal to provide answers of interest for the query *torrhosta* 'dry cough' which is a more specific concept than *hosta* 'cough'. This possibly contradicts seekers' expectation of more detailed queries to result in interesting answers to have a lower rank. Another interesting aspect is the medical versus the layman view on the notion of abortion, where miscarriage in medical terms is an example of abortion. Hence, even though semantic relations found in sources like the UMLS may provide support to portals on deciding their moves, it is important to note that these relations may not reflect the layman view. Consequently, a seeker posting the query *abort* 'abortion' may not be interested in answers related to the more specific query *missfall* 'miscarriage' or 'spontaneous abortion'.

10.2.2.3 Summary

The text level has many opportunities, using semantic resources such as the UMLS, to provide further insights into queries and their relations, but at the same time sources like Snomed CT and MeSH are not able to recognise more than a minor part of all queries. Hence, a semantics-based stylistics analysis will not pay off in a major way until the sources reflect the language actually used by information seekers. Still, in cases when queries are mappable to the UMLS, it reveals interesting aspects like the type of queries posted, especially in cases an inadequate treatment of queries with respect to semantic properties like synonymy and hypernomy.[71] This topic was also addressed in section 9.5 on important properties of a portal, and we may say that in cases with many answers per query it is not only important to present the "right" ones, but the portal moves should respect the semantic relations among queries to ensure a trustworthy portal behaviour.

[71] The use of semantic annotations to gain insights into the use of layman language in relation to established (professional) terminologies and semantic relations among query concepts has also been studied by us in (Oelke et al. 2012).

Query	Interactions	Concept	Related concept	Query	Interactions
influensa	962 (39.9%)	Influenza	Communicable Diseases	infektion +	21 (47.6%)
halsfluss	558 (26.2%)	Tonsillitis	Inflammation	inflammation +	26 (50.0%)
feber	422 (16.6%)	Fever	Body Temperature	kroppstemperatur +	10 (100%)
			Prolonged fever	långvarig feber -	4 (0.0%)
			Febrile Convulsions	feberkramp -	11 (18.2%)
hosta	318 (51.9%)	Coughing	Dry cough	torrhosta -	52 (76.9%)
abort	277 (24.5%)	Unspecified Abortion	Spontaneous abortion	missfall -	170 (40.0%)
			Ectopic Pregnancy	utomkvedshavandeskap -	17 (0.0%)

Table 10.17: Examples of queries with semantically related queries for Västra Götaland, with + indicating a hypernym and - a hyponym. Proportion of clicks with rank>5 in parentheses.

10.3 Answer stylistics

An answer is the result of a portal's move in response to a seeker's utterance. In practice, the answer is a hyperlink, or pointer, to a portal page considered by the portal to provide valuable information given the seeker's move. Since the portal knows it is not able to provide only one answer comprising the one "wanted" by the seeker, it presents several answers decided and sorted according to its preference relation. The choice of answer by the seeker is then registered as a click, and these are provided as part of a dataset, such as the one for Västra Götaland. Thereby, it offers an opportunity to characterise clicks in a similar way as queries, and in section 10.4 this together with the one of the queries allow us to study the query–answer relations, hence the actual search rounds and their stylistics.

An analysis of clicks can be carried out in a similar way as in the case of queries at a substance and text level (section 6.6), where the hyperlink would be the basis for a characterisation of its substance resembling the portal organisation of pages, and the intentions of the structure and pages, or their semantics, to be reflected by the text level of an analysis. In section 10.3.1 we describe some aspects of the clicks not already addressed as part of the presentation of the dataset (section 8.2.4.3), and in section 10.3.2 we present a semantic view at the text level.

An answer stylistics analysis focuses on the portal's moves, especially those leading to a user "acceptance" of the move by clicking an answer. Hence, the portal part of the complete search rounds where we are able to speculate about the pre-query, post-query and answer situations of the seeker, and this is the focus of section 10.4.

10.3.1 Substance

The substance level of answers reflects the outcomes as found in the search log for Västra Götaland, where we have the actual query and the click for the ones where a seeker made a choice among the different answers. Figure 42 presents the number of search rounds per click, and figure 43 the distribution of the seeker vocabulary for Västra Götaland.

By the figures, 7.1% of all search rounds end up among ten different answers, table 10.18, and similarly for 4.0% of the seeker vocabulary. Moreover, 6.4% of the answers are chosen less than ten times, table 10.19, and 36% of the seeker vocabulary. Hence, it is clearly not the case that the search rounds and used vocabulary are evenly spread among the available answers. It is also important to emphasise that we do not know the proportion of the portal vo-

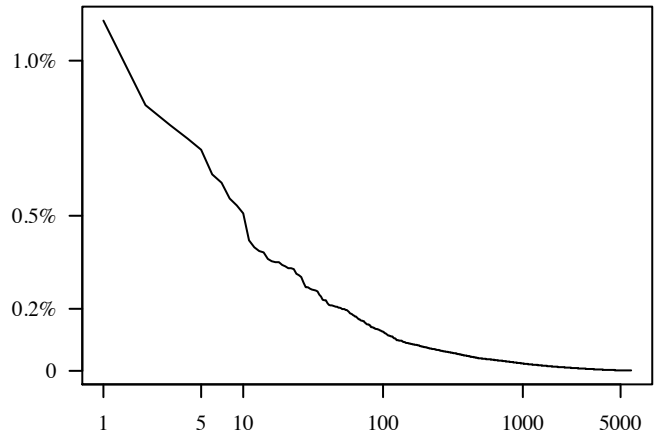


Figure 42: Distribution of search rounds per click for Västra Götaland.

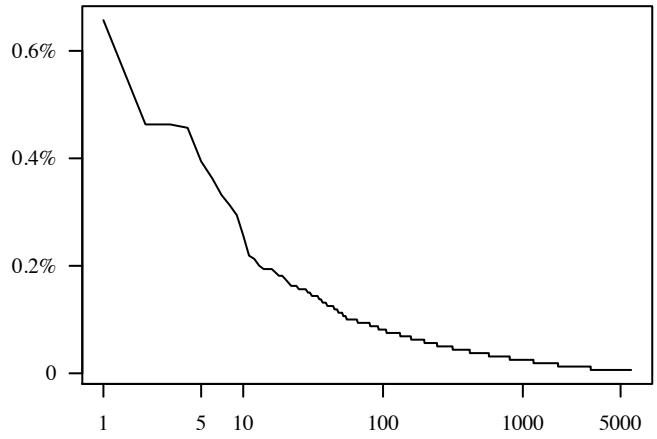


Figure 43: Distribution of the seeker vocabulary per click for Västra Götaland.

cabulary that is actually used by information seekers. One way to characterise the used portal vocabulary is to describe the number of different alternatives per answer level and number of search rounds per alternative, figure 44. For instance, there are 443 different answers related to inquiries with 1836 different queries, with 413 being 6-queries, on advice regarding diseases. Hence, in an average 4.1 queries per answer and 22.5% of the queries being ones requiring more than five answers to be reviewed by a seeker with an interest in this topic. This type of analysis of queries and answers is the focus of the rest of the

chapter, and we will look at some examples of scenarios of importance from a trust perspective for portal providers.

Answer	Clicks
sv/Mina-vardkontakter/	754
sv/vgprimarvard/Fragor-och-svar-om-VG-Primarvard/	572
sv/Regler-och-rattigheter/Resa-till-och-fran-varden/	529
sv/Regler-och-rattigheter/Avgifter-och-ersattningar/Patientfakturering/	500
sv/Mina-vardkontakter/Tjansten-Mina-Vardkontakter-/	476

Table 10.18: Most common clicks for Västra Götaland.

Answer	Clicks
.se/sv/Angereds-narsjukhus1/Angereds-Narsjukhus/	1
%20%c3%a4lsvborg&scoped=true&filter=scope:VGRegionvardportalen#	1
ch.xhtml?q=avgifter%20vid%20akutinl%c3%a4ggning%20p	1
ch.xhtml?q=ambulans&scoped=all&nofilter=scope:VGRegionvardportalen	1
ch.xhtml?q=Azitromax&scoped=all&nofilter=scope:VGRegionvardportalen	1

Table 10.19: Examples of rarely clicked answers for Västra Götaland.

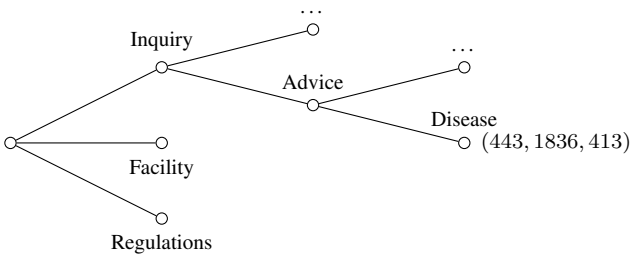


Figure 44: Number of answer alternatives (443), number of queries (1,836) and 6-queries (413) for Västra Götaland.

10.3.2 Text

Table 10.20 shows that almost 30% of all search rounds are related to inquiries on diseases, e.g. facts, self-care and prevention, and in in addition, with over 15% related to advice on symptoms and findings, almost half of the search rounds deal with non-managerial aspects of health care. Moreover, more than one out of ten rounds consider details on health care facilities (15.8%) and

themes like family matters or pregnancy (10.9%). The table also shows that almost 25% of all search rounds for Västra Götaland result on average in more than five answer gists to be read before a decision is made to click one. Moreover, it shows that inquiries related to management of health care activities (39.1%), e.g. local aspects and general support issues, and ones indicating an interest in specific facilities (40.4%) are especially prevalent among the 6-queries. On the other hand inquiries regarding different diseases (17.1%) and themes as stroke care and pregnancy are less common in this category of queries.

Answer		**Search rounds**	**6-queries**	
Level 1	Level 2		Search rounds	Proportion
Inquiry	Disease	19,724 (29.5%)	3,374 (20.3%)	17.1%
Inquiry	Advice	10,859 (16.3%)	2,611 (15.7%)	24.0%
Facility	Location	10,519 (15.8%)	4,250 (25.6%)	40.4%
Inquiry	Theme	7,257 (10.9%)	1,440 (8.7%)	19.8%
Inquiry	Regulation	5,653 (8.5%)	1,140 (6.9%)	20.2%
Inquiry	Management	5,311 (8.0%)	2,075 (12.5%)	39.1%
Inquiry	Procedure	4,178 (6.3%)	1,034 (6.2%)	24.8%
Inquiry	News	1,773 (2.7%)	473 (2.9%)	26.7%
Inquiry	Support	698 (1.0%)	193 (1.2%)	27.7%
Inquiry	–	506 (0.8%)	11 (0.1%)	2.2%
Inquiry	Group (newborn)	211 (0.3%)	6 (0.04%)	2.8%
–	–	12 (0.01%)	3 (0.02%)	25.0%
Regulation	Organisation	10 (0.01%)	0	0.0%
		66,783 (100%)	16,616 (100%)	

Table 10.20: Search rounds per answer level 2 for Västra Götaland, including proportion of search rounds containing 6-queries.

If we further divide the answers in sub-categories, table 10.21 shows that 6-queries seem to be overrepresented among managerial topics related to local aspects and general support. Similarly, it highlights fewer disease inquiries related to self-care and facts among these. Moreover, almost 25% of all Västra Götaland search rounds are related to inquiries on facts about diseases by queries as *influensa* 'influenza'.

Another large group of rounds is (implicitly) related to locations, e.g. *egenremiss* 'self-referral' and *mammografi* 'mammography'. These two seeker interests make up more than a third of all search rounds, but in the first case these are rather seldom (14.9%) a result of a 6-query and in the latter case rather often (40.4%) the result of a 6-query. It is also interesting to note that inquiries

Answer				Search rounds	6-queries	
Level 1	Level 2	Level 3	Level 4		Search rounds	Proportion
Inquiry	Disease	Disease	Sjukvardsradgivningen	15,542 (23.3%)	2,311 (13.9%)	14.9%
Facility	Location	–	–	10,519 (15.8%)	4,250 (25.6%)	40.4%
Inquiry	Procedure	Drug	Sjukvardsradgivningen	2,121 (3.2%)	407 (2.5%)	19.2%
Inquiry	Theme	Group (family)	Sjukvardsradgivningen	2,093 (3.1%)	392 (2.4%)	18.7%
Inquiry	Management	Location	–	1,468 (2.2%)	1,147 (6.9%)	78.1%
Inquiry	Management	–	–	1,434 (2.1%)	405 (2.4%)	28.2%
Inquiry	News	–	–	1,121 (1.7%)	184 (1.1%)	16.4%
Inquiry	Management	Support	–	1,094 (1.6%)	443 (2.7%)	40.5%
Inquiry	Disease	Self-care	Sjukvardsradgivningen	877 (1.3%)	134 (0.8%)	15.3%
Inquiry	Procedure	Procedure (therapeutic)	Sjukvardsradgivningen	792 (1.2%)	151 (0.9%)	19.1%
Inquiry	Procedure	Procedure (diagnostic)	Sjukvardsradgivningen	787 (1.2%)	277 (1.7%)	35.2%
Inquiry	Regulation	Delivery	Vardgaranti	684 (1.0%)	401 (2.4%)	58.6%
Inquiry	Theme	Travel	Sjukvardsradgivningen	642 (1.0%)	254 (1.5%)	39.6%
				39,174 (57.5%)	10,756 (64.8%)	

Table 10.21: Common answers for Västra Götaland, including proportion of search rounds containing 6-queries.

on procedures do not stand out, but with slightly larger proportion rounds with 6-queries regarding diagnostics in comparison to therapeutics.

Considering disease-related rounds, table 10.22, the ones related to non-stroke patient groups or prevention are more prevalent among rounds with clicks on higher rank, but ones related to stroke more often need less than six answers to satisfy the seeker.

Turning our interest to advice seeking inquiries which are mainly related to diseases and care providers, table 10.23, the interests in stomach ache and fever stand out among the queries, with the first type of inquiries being more prevalent among the high-rank clicks and the latter being underrepresented in the group.

The categories of pages, or answer types, presented above differ from the thematic ones, where many different aspects like diseases, care providers and advice are organised around a theme like family, travel or elderly. Based on table 10.24, especially topics related to family matters are not as well represented among 6-queries, but ones related to travel and summer issues like holidays and sunburn are more common among these inquiries.

Considering management topics, table 10.25 confirms location and general support interests to more often lead to higher than average rank of answers of interest, and topics related to administrative questions like rescheduling an appointment, changing care providers and regional health information to more often be of lower rank.

To summarise, table 10.26, in general it seems like queries originating in an interest in *health care units* tend to result in more challenges than ones related to other types of inquiries. When the interest targets *prevention* or specific *patient groups*, these also more often result from 6-queries. However, considering procedures like *diagnostics* and *therapeutics* only the former shows a slight tendency towards 6-queries, and when seeking advice on signs or symptoms the results are inconclusive with stomach ache answers to clearly originate in 6-queries, but fever-related information easier to find. Similar things are also found for thematic answers.

To fully understand these findings, we will in the next section study them in detail in relation to the queries.

Answer			Search rounds	6-queries	
Level 3	Level 4	Level 4 Concept		Search rounds	Proportion
Disease	Health Care Activity	Advice	15,542 (23.3%)	2,311 (13.9%)	14.9%
Location	–	–	1,306 (2.0%)	458 (2.8%)	35.1%
Self-care	Health Care Activity	Advice	877 (1.3%)	134 (0.8%)	15.3%
Disease	–	–	177 (0.3%)	17 (0.1%)	9.6%
Group (patient)	Disease or Syndrome	Cerebrovascular accident	156 (0.2%)	10 (0.1%)	6.4%
Procedure (preventive)	–	–	149 (0.2%)	99 (0.6%)	66.4%
Self-care	–	–	138 (0.2%)	2 (0.01%)	1.5%
Group (patient)	Neoplastic Process	Malignant neoplasm of breast	90 (0.1%)	64 (0.4%)	71.1%
Group (patient)	Bacterium	Borrelia	86 (0.1%)	46 (0.3%)	53.5%
Group (patient)	Disease or Syndrome	Amyotrophic Lateral Sclerosis	80 (0.1%)	0	0.0%
Group (patient)	Disease or Syndrome	Irritable Bowel Syndrome	74 (0.1%)	34 (0.2%)	45.9%
Group (patient)	Disease or Syndrome	Ulcerative Colitis	56 (0.1%)	38 (0.2%)	67.9%
Group (patient)	Disease or Syndrome	Rheumatism	53 (0.1%)	37 (0.2%)	69.8%
Drug	–	–	22 (0.03%)	22 (0.1%)	100%
			18,806 (28.1%)	3,272 (19.7%)	

Table 10.22: Answers related to disease for Västra Götaland, including proportion of search rounds containing 6-queries.

Answer			Search rounds	6-queries	
Level 3	Level 4	Level 4 Concept		Search rounds	Proportion
Location	–	–	591 (0.9%)	156 (0.9%)	26.4%
Disease	Health Care Activity	Self-care interventions	374 (0.6%)	49 (0.3%)	13.1%
Disease	Finding	Fever	227 (0.3%)	8 (0.1%)	3.5%
Disease	Sign or Symptom	Stomach ache	221 (0.3%)	74 (0.5%)	33.5%
Disease	Sign or Symptom	Diarrhea and vomiting, symptom	171 (0.3%)	28 (0.2%)	16.4%
Disease	Disease or Syndrome	Urinary tract infection	148 (0.2%)	12 (0.1%)	8.1%
Disease	Disease or Syndrome	Tonsillitis	135 (0.2%)	17 (0.1%)	12.6%
Drug	Pharmacologic Substance	Contraceptive Agents	115 (0.2%)	14 (0.1%)	12.2%
Group (patient)	–	–	105 (0.2%)	19 (0.1%)	18.1%
Disease	Disease or Syndrome	Herpes zoster disease	94 (0.1%)	6 (0.04%)	6.4%
Disease	Disease or Syndrome	Borrelia Infections	85 (0.1%)	53 (0.3%)	62.4%
Disease	Disease or Syndrome	Other vitamin B12 deficiency anemias	81 (0.1%)	51 (0.3%)	63.0%
Group (patient)	Bacterium	Borrelia	79 (0.1%)	38 (0.2%)	48.1%
Disease	Finding	Mass in breast	64 (0.1%)	51 (0.3%)	79.7%
Disease	Sign or Symptom	Dyspepsia	60 (0.1%)	55 (0.3%)	91.7%
Disease	Disease or Syndrome	Leg Ulcer	58 (0.1%)	52 (0.3%)	89.7%
Procedure (diagnostic)	Health Care Activity	Prenatal care	47 (0.1%)	39 (0.2%)	83.0%
			2,655 (4.0%)	722 (4.7%)	

Table 10.23: Answers related to advice seeking for Västra Götaland, including proportion of search rounds containing 6-queries.

Answer	**Search rounds**	**6-queries**	
Level 3		Search rounds	Proportion
Group (family)	2,153 (3.2%)	404 (2.4%)	18.8%
Pregnancy	1,807 (2.7%)	369 (2.2%)	20.4%
Travel	704 (1.1%)	256 (1.5%)	36.4%
Procedure	562 (0.8%)	36 (0.2%)	6.4%
Group (elderly)	501 (0.8%)	86 (0.5%)	17.2%
Summer	443 (0.7%)	135 (0.8%)	30.5%
News	412 (0.6%)	73 (0.4%)	17.7%
Health	362 (0.5%)	27 (0.2%)	7.5%
Heart	241 (0.4%)	45 (0.3%)	18.7%
Cancer	66 (0.1%)	9 (0.1%)	13.6%
Total:	7,251 (10.9%)	1,440 (8.6%)	

Table 10.24: Common thematic answers for Västra Götaland, including proportion of search rounds containing 6-queries.

Answer	**Search rounds**	**6-queries**	
Level 3		Search rounds	Proportion
Location	1,468 (2.2%)	1,147 (6.9%)	78.1%
Support	1,094 (1.6%)	443 (2.7%)	40.4%
Administration	590 (0.9%)	49 (2.9%)	8.3%
Information	572 (0.9%)	8 (0.05%)	1.4%
Language	123 (0.2%)	18 (0.1%)	14.6%
Regulation	30 (0.04%)	5 (0.03%)	16.7%

Table 10.25: Answers related to health care management for Västra Götaland, including proportion of search rounds containing 6-queries.

6-queries	**Non-6-queries**
health care units	theme (family)
prevention	patient group (stroke)
patient groups (MS)	symptom (fever)
diagnostics	
theme (travel)	
symptom (stomach ache)	

Table 10.26: Origin of common answer types for Västra Götaland.

10.4 Interaction stylistics

Depending on the sophistication of the portal, the link between query and answer may be found on a spectrum from simple lookup of the query at the portal page, which can be called a *substance matching*, to an attempt by the portal to decode the intentions of the seeker, e.g. an interest in treatments when symptom queries are posted, called *discourse mapping*. In between these we find approaches taking into consideration more or less lexical information, e.g. deciding the semantics of the query using available lexicon and terminology information. Thereby, this last level of query analysis could be called *lexico-semantic mapping*.

Table 10.27 clearly shows that there are a number of types of interactions which stand out in Västra Götaland. First of all, an interest in management of *mina vårdkontakter* 'my care contacts' represents 3.0% of all interactions, and in half of the cases this interest results in more than five answers to be considered before finding one of potential interest. Two similar types of common interactions are ones related to *egenremiss* 'self-referral' (3.0%) and *sjukresor* 'patient transports' (1.5%). In the first case, an interest in self-referral is tied to specific facilities offering it, possibly for specific diseases, and regulation of its management. The interest in regulation is shared by queries on patient transport, and in this case often in connection with support on how it works. Both these types of interactions indicate that for terms related to the management of one's own care, e.g. *mina vårdkontakter*, there are many questions on its functions and regulations. If we study these three queries in detail, we have that their distribution among types of answers, table 10.28, shows that this type of interests often boils down to a number of areas of concern and one could consider all of them to be covered in fact sheets indexed by corresponding queries and always presented as first choice. It is also important to note that challenges to find the interesting information differs substantially both within and between the areas, with regulatory aspects in some cases never leading to any problems to find interesting answers among the first suggestions and in other cases lead substantial time spent to find an answer.

From a search game perspective, we have a group of seeker moves with a narrow set of satisfying portal moves, and either the portal would try to decide the best for each seeker move or one move ending up in an answer addressing all needs. An open question is if the seeker would consider the first or second approach most trustworthy. In the first case the answer of interest would end with a high rank, and in the second case the answer may be considered too general or too "long". Hence, not a trivial choice, but the point we want to make is that the analysis reveals two options for further studies.

We have a similar finding in the case of *mammografi* 'mammography', ta-

Query	**Answer**				**Query–answer**	**Rank>5**
	Level 1	Level 2	Level 3	Level 4		
mina vårdkontakter	Inquiry	Management	–	–	1752	52.7%
egenremiss	Facility	–	–	–	1392	83.3%
influensa	Inquiry	Disease	Disease	Advice	747	32.7%
vårdgaranti	Inquiry	Regulation	Delivery	–	562	59.3%
sjukresor	Inquiry	Regulation	Transport	–	530	58.7%
mammografi	Facility	–	–	–	528	64.2%
urinvägsinfektion	Inquiry	Disease	Disease	Advice	412	17.5%
gravid	Inquiry	Theme	Pregnancy	–	382	40.3%
klamydia	Inquiry	Disease	Disease	Advice	335	60.0%
njursten	Inquiry	Disease	Disease	Advice	258	7.0%
egenremiss	Inquiry	Regulation	Delivery	–	243	0.0%
magkatarr	Inquiry	Disease	Disease	Advice	241	28.2%
högkostnadsskydd	Inquiry	Regulation	Fees	–	239	51.5%
ibs	Inquiry	Disease	Disease	Advice	223	44.8%
halsfluss	Inquiry	Theme	Group (family)	Advice	222	24.3%
egenremiss	Inquiry	Disease	Location	–	220	0.0%
bensår	Inquiry	Disease	Disease	Advice	196	44.4%
sjukresor	Inquiry	Support	Transport	–	195	39.5%
sjukresor	Inquiry	Regulation	–	–	183	0.0%
feber	Inquiry	Advice	Disease	Fever	181	0.0%

Table 10.27: Most common query–answer pairs for Västra Götaland.

Query	**Answer**				**Query–answer**	**Rank>5**
	Level 1	Level 2	Level 3	Level 4		
mina vårdkontakter	Inquiry	Management	–	–	1752	52.7%
egenremiss	Facility	–	–	–	1392	83.3%
sjukresor	Inquiry	Regulation	Transport	–	530	58.7%
egenremiss	Inquiry	Regulation	Delivery	–	243	0.0%
egenremiss	Inquiry	Disease	Location	–	220	0.0%
sjukresor	Inquiry	Support	Transport	–	195	39.5%
sjukresor	Inquiry	Regulation	–	–	183	0.0%
egenremiss	Inquiry	Procedure	Procedure (diagnostic)	Hereditary Diseases	165	0.0%
mina vårdkontakter	Inquiry	Management	Support	–	115	100%
sjukresor	Inquiry	Regulation	Fees	–	115	0.0%
mina vårdkontakter	Inquiry	News	–	–	112	0.0%
mina vårdkontakter	Facility	–	–	–	17	0.0%

Table 10.28: Distribution of answers to queries regarding health care management for Västra Götaland.

ble 10.29, where the query clearly indicates an interest in specific care givers, possibly for management of appointments. The problem of finding interesting answers when the focus is on facilities, as also holds for queries related to self-referral, could be explained by the portal treating each facility as a separate "answer" and thereby there is a risk of the one of interest not being the one promoted by the portal. However, this also raises a question on how context information on seeker location is, and should, be used in the choice of portal moves.

If we turn our interest to the queries which are automatically mappable to the UMLS, hence possible to analyse at a text level and where semantic information may be used in the selection of portal moves, we find interactions originating in different types of diseases to make up a major part of the most common ones, except for mammography as a procedure related to an interest in health care units and interest in symptoms like fever, table 10.30. From a 6-queries perspective, we have already seen the problems whenever there is an interest in facilities to find an adequate answer in five, or less, answers, but it is interesting to note rather well-defined concepts like tonsillitis and irritable bowel syndrome to be among those where in more than 40% of the cases a query will result in clicks of rank higher than five, at the same time as an interest in kidney calculi or pneumonia, and even symptoms like fever, do not result in any seeker challenges. The differences in treatment of disease-related queries, with some being 6-queries and others never being in need of longer answers sets raises a question of how seekers will view the competence of portal, thereby if and when to trust the portal.

The interest in interactions related to diseases and symptoms is further supported by a semantic analysis of the search rounds, table 10.31, revealing 27% of all interactions to begin with queries with disease concepts and answers to consist of either disease facts or advice. It is also clear that the existence of thematic portal areas related to, for instance, families, pregnancy and patient groups, impact the interactions. In a search game context, this can be seen as, unintentional, attempts to guide the seekers in certain directions, cf the example on stroke and fatigue in section 5.2.1.

As seen in several earlier cases, there is a general problem in cases when seekers are interested in health care units related to, for instance, procedures like *mammografi* 'mammography' or diseases as *sle* (Systemic Lupus Erythematosus). However, in 1177 Vårdguiden the support for identifying relevant health care units has been improved, but even for the studied time periods it was possible for the portals to make use of location information on seekers to constrain the answers (section 9.2.2). Still the search logs do not support any hypothesis that this has taken place, or led to improved usability considering searches with interest in location-dependent information. The topic of

Answer				Query–answer	Rank>5
Level 1	Level 2	Level 3	Level 4		
Facility	–	–	–	528	64.2%
Inquiry	Procedure	Procedure (diagnostic)	Advice	120	42.5%
Inquiry	Disease	Location	–	107	8.4%
Inquiry	Management	–	–	44	100%
Inquiry	Advice	Disease	Mass in breast	39	100%
Inquiry	Theme	News	Malignant neoplasm of breast	32	0.0%
Inquiry	Management	Support	–	32	100%
Inquiry	Advice	Location	–	26	100%
Inquiry	Advice	Disease	Malignant neoplasm of breast	17	100%
Inquiry	Advice	Procedure (diagnostic)	Mammography	17	100%

Table 10.29: Distribution of answers to queries regarding mammography for Västra Götaland.

Concept	**Answer**				**Concept–answer**	**Rank>5**
	Level 1	Level 2	Level 3	Level 4		
Influenza	Inquiry	Disease	Disease	Advice	747	32.7%
Mammography	Facility	–	–	–	528	64.2%
Patient currently pregnant	Inquiry	Theme	Pregnancy	–	382	40.3%
Gastritis	Inquiry	Disease	Disease	Advice	307	22.1%
Kidney Calculi	Inquiry	Disease	Disease	Advice	258	7.0%
Herpes zoster disease	Inquiry	Disease	Disease	Advice	249	10.0%
Irritable Bowel Syndrome	Inquiry	Disease	Disease	Advice	223	44.8%
Tonsillitis	Inquiry	Theme	Group (family)	Advice	222	24.3%
Leg Ulcer	Inquiry	Disease	Disease	Advice	196	44.4%
Fever	Inquiry	Advice	Disease	Fever	181	0.0%
Cholecystolithiasis	Inquiry	Disease	Disease	Advice	166	19.3%
Pneumonia	Inquiry	Disease	Disease	Advice	162	1.9%
Tonsillitis	Inquiry	Disease	Disease	Advice	151	47.0%
Brain Concussion	Inquiry	Disease	Disease	Advice	149	16.8%
Cerebrovascular accident	Inquiry	Disease	Group (patient)	Cerebrovascular accident	137	4.4%
Condyloma	Inquiry	Disease	Disease	Advice	133	18.8%
Unspecified Abortion	Inquiry	Theme	Pregnancy	–	133	0.0%
Hepatitis C	Inquiry	Disease	Disease	Advice	132	40.2%
Fever	Inquiry	Disease	Disease	Advice	127	9.4%
Varicosity	Inquiry	Disease	Disease	Advice	127	44.1%

Table 10.30: Most common concept–answer pairs for Västra Götaland.

Semantic type	Answer				Concept–answer	Rank>5
	Level 1	Level 2	Level 3	Level 4		
Disease or Syndrome	Inquiry	Disease	Disease	Health Care Activity	6,211	23.9%
Disease or Syndrome	Inquiry	Advice	Disease	Disease or Syndrome	1,706	13.7%
Disease or Syndrome	Inquiry	Theme	Group (family)	Health Care Activity	922	27.0%
Sign or Symptom	Inquiry	Disease	Disease	Health Care Activity	749	27.6%
Pharmacologic Substance	Inquiry	Procedure	Drug	Health Care Activity	687	14.6%
Diagnostic Procedure	Facility	–	–	–	580	60.9%
Finding	Inquiry	Theme	Pregnancy	–	545	29.2%
Disease or Syndrome	Inquiry	Disease	Group (patient)	Disease or Syndrome	544	14.9%
Finding	Inquiry	Disease	Disease	Health Care Activity	518	26.6%
Organic Chemical	Inquiry	Procedure	Drug	Health Care Activity	507	14.8%
Therapeutic or Preventive Procedure	Facility	–	–	–	494	42.5%
Body Substance	Inquiry	Disease	Disease	Health Care Activity	474	13.7%
Disease or Syndrome	Inquiry	Procedure	Drug	Health Care Activity	418	33.5%
Body Part, Organ or Organ Component	Inquiry	Disease	Disease	Health Care Activity	412	28.9%
Disease or Syndrome	Facility	–	–	–	409	40.6%
Pathologic Function	Inquiry	Disease	Disease	Health Care Activity	338	34.3%
Sign or Symptom	Inquiry	Advice	Disease	Sign or Symptom	326	20.9%
Disease or Syndrome	Inquiry	Advice	Disease	Sign or Symptom	326	31.9%
Disease or Syndrome	Inquiry	Disease	Location	–	313	6.4%
Professional or Occupational Group	Facility	–	–	–	298	25.5%

Table 10.31: Most common semantic type–answer pairs for Västra Götaland.

scenarios where seekers are interested in health care units is further analysed in section 10.4.1.

Table 10.31 also shows that it seems like pharmacology-related queries in many cases lead to no problems finding adequate information, and this topic is discussed in section 10.4.2 with a focus on a portal's treatment of often uniquely defined names of pharmaceutical products. In a similar way we study concepts corresponding to body parts, where there might be a difference between laymen and professional terminology, but also if the interest is in the body part itself or, for instance, the many diseases or symptoms involving it. We end with a study of the possibly most common reason for search a health portal, i.e. symptoms and diseases, in section 10.4.3.

10.4.1 Location-dependent queries

Since our interest is mainly in searches with a tendency to lead to seekers having to review more than five answers, we will study queries where the seeker had an interest in location-related topics, e.g. health care units, especially those originating with queries expressed in terms of *procedures*, *diseases* and *professions*.

By table 10.32 we see that the facility-related answers make up 34% of the ones starting with a procedure query, to be compared to 23% related to the procedures themselves and 10% to thematic answers.

For procedures, it is interesting to note that many of the diagnostic procedures occur in the setting of facilities, table 10.33, and may reveal a scenario of *diagnostics–facility* interactions. However, queries related to *gastroskopi* 'gastroscopy' and *lungröntgen* 'lung x-ray' will in less than 10% of the rounds lead to facility clicks. This is even more interesting, when procedures like endoscopy and colonoscopy in more than 55% of the search rounds end up with answers of interest to consider facilities. If we compare these examples of queries we see a clear difference in the type of interests for these in one way similar, but in another way different, queries, table 10.34. Moreover, if we study therapeutic and preventive procedures, the interest in facility answers is not as clear. This shows that concepts related to procedures often have a two-fold intention, both as the procedure itself and as an indicator of an interest in health care units administrating, or offering, the procedure. From a search game perspective, this analysis shows that for certain types of seeker moves the portal may benefit from providing combined answers on both facts on procedures and their administrators. This is also an interesting approach, since we may hypothesise that the reason for posting a query on a procedure is probably that the seeker is to be involved in the procedure, and thereby visit

Semantic type	Answer			Semantic type–answer	Rank>5
	Level 1	Level 2	Level 3		
Diagnostic Procedure	Facility	–	–	580	60.9%
Therapeutic or Preventive Procedure	Facility	–	–	494	42.5%
Diagnostic Procedure	Inquiry	Procedure	Procedure (diagnostic)	256	29.3%
Therapeutic or Preventive Procedure	Inquiry	Theme	Pregnancy	179	14.5%
Therapeutic or Preventive Procedure	Inquiry	Procedure	Procedure (therapeutic)	138	25.4%
Diagnostic Procedure	Inquiry	Disease	Location	109	8.3%
Therapeutic or Preventive Procedure	Inquiry	Disease	Disease	108	11.1%
Therapeutic or Preventive Procedure	Inquiry	News	Procedure (preventive)	104	71.2%
Diagnostic Procedure	Inquiry	Advice	Disease	88	65.9%
Therapeutic or Preventive Procedure	Inquiry	Disease	Procedure (preventive)	87	21.8%
Therapeutic or Preventive Procedure	Inquiry	Theme	Group (family)	70	14.3%
Therapeutic or Preventive Procedure	Inquiry	Procedure	Drug	67	3.0%
Therapeutic or Preventive Procedure	Inquiry	Theme	Travel	60	86.7%
Therapeutic or Preventive Procedure	Inquiry	Disease	Group (patient)	60	1.7%
Therapeutic or Preventive Procedure	Inquiry	Advice	Procedure (therapeutic)	55	23.6%
Therapeutic or Preventive Procedure	Inquiry	Advice	Location	51	35.3%
Diagnostic Procedure	Inquiry	Management	–	48	91.7%
Diagnostic Procedure	Inquiry	Advice	Procedure (diagnostic)	44	56.8%
Therapeutic or Preventive Procedure	Inquiry	Procedure	Procedure (diagnostic)	44	13.6%
Therapeutic or Preventive Procedure	Inquiry	Advice	Disease	43	44.2%

Table 10.32: Common answers to procedure-related queries for Västra Götaland.

Query	Search rounds	Proportion answer
mammografi	962	54.9%
endoskopi	23	60.9
koloskopi	18	55.6%
arbets-ekg	17	94.1%
gastroskopi	14	7.1%
lungröntgen	12	8.3%
ctc	9	77.8%
angiografi	1	100%
pet	1	100%
vc	1	100%

Table 10.33: Common diagnostic procedure queries with facility clicks, with proportion of queries with the given click type.

Answer type	Gastroscopy	Endoscopy	Colonoscopy
Inquiry/Procedure/Procedure (diagnostic)	9	–	2
Inquiry/Advice/Procedure (diagnostic)	4	–	6
Facility/Location	1	14	10
Inquiry/Disease/Disease	–	5	–
Inquiry/Management/Location	–	4	–

Table 10.34: Distribution of answers for queries related to gastro-, endo- and colonoscopy for Västra Götaland.

a facility. Hence, a combined answer may be viewed by the seeker as not only trustworthy, but even insightful.

If we study disease-related queries, table 10.35, we see that only queries regarding *könssjukdomar* 'sexually transmitted diseases' and *anorexia* are more prevalent in the setting of location answers, and a reason may be that these do often have dedicated care units. However, a scenario on diseases and facilities needs to be defined at query level and not the level of semantic types as in the previous examples. This is also supported by table 10.36, where we see that facility in the case of diseases is only a minor interest.

Finally, if we consider professions, tables 10.37 and 10.38, it is clear that more than half of all profession-related interactions reveal a facility interest. However, the query kiropraktor 'chiropractor' does not follow this pattern, which might be due to this profession being less well-known than, for instance, gynaecologist or psychologist, and before looking for a care provider the seeker wants to know more about the discipline as in the case of some procedures.

The study of location-dependent queries not only shows that certain types of queries are of higher probability to be related to interest in care providers, but also a general opportunity for portals to utilise semantic information of seekers' moves, but also of past rounds, to identify (semantic) scenarios to guide the portals in their decision on moves to better satisfy the seekers' needs. However, as exemplified by the disease queries, in some cases the decision will have to be made at substance level, and not at text level, or as in the case of procedures, the portal may benefit from moves which are combined of several potential answer types.

Query	**Search rounds**	**Proportion answer**
stroke	339	12.7%
njursten	307	0.7%
diabetes	270	10.0%
hepatit c	209	12.4%
artros	162	8.0%
astma	157	17.2%
gikt	148	14.2%
tbe	148	24.3%
sle	136	6.6%
endometrios	124	2.4%
fetma	117	29.9%
fibromyalgi	88	11.4%
migrän	66	10.6%
njursvikt	62	1.6%
blåskatarr	58	12.1%
epilepsi	50	8.0%
anorexia	37	45.9%
könssjukdomar	32	81.3%
add	24	20.8%
infektion	21	33.3%

Table 10.35: Common disease queries with facility clicks, with proportion of queries with the given click type.

Answer			Semantic type–answer	Rank>5
Level 1	Level 2	Level 3		
Inquiry	Disease	Disease	6255	23.8%
Inquiry	Advice	Disease	2445	20.0%
Inquiry	Theme	Group (family)	931	26.7%
Inquiry	Disease	Group (patient)	584	13.9%
Inquiry	Procedure	Drug	421	33.7%
Facility	–	–	409	40.6%
Inquiry	Disease	Location	313	17.3%
Inquiry	Disease	Self-care	306	12.1%
Inquiry	Theme	Travel	219	27.3%
Inquiry	Theme	Pregnancy	192	27.1%
Inquiry	Advice	Group (patient)	168	25.6%
Inquiry	News	Procedure (preventive)	166	0.0%
Inquiry	News	–	137	35.0%
Inquiry	Advice	Location	117	17.1%
Inquiry	Theme	News	109	48.6%
Inquiry	Procedure	Procedure (therapeutic)	91	12.1%
–	–	–	77	0.0%
Inquiry	Advice	Drug	73	30.1%
Inquiry	Advice	Procedure (diagnostic)	62	80.6%
Inquiry	Procedure	Procedure (diagnostic)	53	58.5%

Table 10.36: Common answers to disease-related queries for Västra Götaland.

Query	**Search rounds**	**Proportion answer**
kiropraktor	138	30.4%
gynekolog	90	100%
ortoped	55	85.5%
dietist	44	79.5%
psykolog	24	100%
ögonläkare	14	78.6%
urolog	14	50.0%
audionom	12	91.7%
ortoptist	12	33.3%
tolk	9	55.6%
arbetsterapeut	8	75.0%
neonatal	4	100%
sjukgymnast	4	50.0%
läkare	3	100%
psykiatriker	3	100%
specialisttandläkare	2	100%
diabetessköterska	1	100%
neurolog	1	100%

Table 10.37: Common profession and occupation queries with facility clicks, with proportion of queries with the given click type.

Answer			Semantic type–answer	Rank>5
Level 1	Level 2	Level 3		
Facility	–	–	298	25.5%
Inquiry	Advice	Disease	77	79.2%
Inquiry	Procedure	Procedure (therapeutic)	26	0.0%
Inquiry	Theme	Procedure	22	22.7%
Inquiry	Disease	Disease	19	57.9%
Inquiry	Advice	Procedure (therapeutic)	15	100%
Inquiry	Theme	Health	11	0.0%
Inquiry	Regulation	Support	9	11.1%
Inquiry	Regulation	Laws	7	0.0%
Inquiry	Theme	Group (family)	6	33.3%
Inquiry	Disease	Group (patient)	5	60.0%
Inquiry	Disease	Location	5	0.0%
Inquiry	News	–	2	0.0%
Inquiry	Regulation	Fees	1	0.0%
–	–	–	1	0.0%
Inquiry	Advice	Location	1	0.0%

Table 10.38: Common answers to profession-related queries for Västra Götaland.

10.4.2 Queries with named entities

As seen by tables 10.39 and 10.40, people searching for information related to pharmacologic substances often do so using specific product names, and with an interest in the product itself or an (intended) disease for the treatment. In these cases, product specific portals like FASS (LIF 2013) would be an adequate complement to a potentially more symptom/disease centric portal. Or in terms of portal moves, if different (sub-)portals, and vocabularies, may be used for different types of questions, thereby also with a potentially different decision procedure for portal moves.

The question on treatment of queries with drug names highlights the importance of portal management of named entities in general, possibly posted as acronyms or by synonyms. From a seeker perspective, this is an important topic, since the used seeker vocabulary probably consists of terms in a language not well-known to her. This may result in problems on how to proceed if the portal runs into challenges to provide adequate answers.[72]

[72]The use of named entities like drugs has been studied by us in (Eklund and Kokkinakis 2012).

Query	Search rounds	Proportion answer
waran	106	41.5%
preventivmedel	101	25.7%
acne	62	32.3%
folsyra	62	41.9%
tramadol	55	92.7%
alkohol	47	8.5%
p-piller	40	15.0%
tryptizol	38	5.3%
prednisolon	33	90.9%
atacand	25	64.0%
citodon	25	36.0%
kortison	22	72.7%
simvastatin	21	42.9%
fragmin	19	21.1%
viagra	19	68.4%
atarax	18	66.7%
flagyl	17	88.2%
provera	17	70.6%
insulin	16	87.5%
cerazette	15	60.0%

Table 10.39: Common pharmacology queries, with proportion of the clicks related to pharmaceuticals.

Answer			Search	Rank>5	Rank>5
Level 1	Level 2	Level 3	**rounds**		
Inquiry	Procedure	Drug	502	84	**16.7%**
Inquiry	Disease	Disease	164	69	42.1%
Inquiry	Advice	Disease	157	42	26.8%
Inquiry	Advice	Drug	154	27	**17.5%**
Facility	Location	–	133	88	66.2%
Inquiry	Theme	Pregnancy	77	3	3.9%
Inquiry	Disease	Group (patient)	58	0	0%
Inquiry	Theme	Group (elderly)	56	2	3.6%
Inquiry	Theme	News	38	10	26.3%
Inquiry	Advice	Group (patient)	33	2	6.1%
Inquiry	Procedure	Procedure (therapeutic)	26	1	3.8%
Inquiry	Theme	Group (family)	23	16	69.6%
Inquiry	News	–	20	7	35.0%
Inquiry	Theme	Procedure	20	0	0%
Inquiry	Advice	Location	15	2	13.3%
Inquiry	Theme	Travel	14	10	71.4%
Inquiry	Regulation	Fees	12	0	0%
Inquiry	Disease	Location	11	2	18.2%
Inquiry	Advice	Procedure (diagnostic)	8	5	62.5%
Inquiry	Theme	Heart	6	1	16.7%

Table 10.40: Answers to queries of semantic type Pharmacologic Substance for Västra Götaland.

10.4.3 Queries with diverse answer sets

In the previous section, the focus was often on the concept itself, but in the case of body parts, table 10.41 shows that the interest seems to be more diverse. For instance, in almost every case the query *blindtarm* 'appendix' results in answers on facts or advice in relation to the disease. However, for the query *bröst* 'breast' this interest is less than 7%. Hence, it is difficult to see why some diseases are related to different topics, but it is worth noting that queries like *bröst* are also covered by thematic answers on pregnancy and news issued by the portal providers.

One of the most common types of (mappable) queries are the ones representing signs or symptoms, but as shown by table 10.42, it is very difficult to see clear patterns of why some symptoms indicate an interest in facts

Query	Search rounds	Facts	Advice	Procedure	Location
njursten	307	84.0%	15.3%	0.0%	0.7%
gallsten	229	72.5%	17.5%	6.1%	0.0%
blindtarm	106	81.1%	18.9%	0.0%	0.0%
ögon	96	49.0%	0.0%	6.3%	26.0%
sköldkörtel	89	59.6%	21.3%	9.0%	10.1%
bröst	58	6.9%	0.0%	0.0%	15.5%
knä	57	17.5%	14.0%	43.9%	0.0%
urin	37	78.4%	0.0%	21.6%	0.0%
lymfkörtlar	35	85.7%	2.9%	0.0%	0.0%
aorta	32	15.6%	0.0%	0.0%	0.0%
ganglion	28	100%	0.0%	0.0%	0.0%
revben	28	78.6%	21.4%	0.0%	0.0%
hälsena	26	30.8%	34.6%	0.0%	0.0%
förhud	25	72.0%	28.0%	0.0%	0.0%
prostata	25	0.0%	0.0%	60.0%	16.0%

Table 10.41: Common body substance or body part queries, with proportion of the clicks related to different aspects on diseases.

on a disease, and others in different types of advice. Two extremes are the queries *mjölkstockning* 'milk stasis' and *dyspepsi* 'dyspepsia'. But we also have queries like *ischias* 'sciatica' and *huvudvärk* 'headache', where the interest tend to be of equal size.

This type of queries where it is often unclear if the interest is in facts or advice, hence two rather different kinds of portal moves, is a challenge for a portal to maintain the seekers' trust.

As seen by table 10.43, this division among different types of interests is also a challenge in the case of disease queries.

To conclude, for some types of queries, e.g. body parts, symptoms and diseases, a seeker's interest is not easily revealed by the query itself.

Query	Search rounds	Facts	Advice	Procedure	Location
hosta	318	33.3%	16.4%	0.0%	0.0%
ischias	256	39.1%	42.2%	0.0%	0.0%
yrsel	164	62.8%	26.8%	0.0%	0.0%
förstoppning	123	61.8	30.1%	0.0%	0.0%
huvudvärk	112	32.1%	36.6%	0.0%	0.0%
diarré	99	19.2%	48.5%	0.0%	0.0%
kärlkramp	86	19.8%	12.8%	0.0%	0.0%
mjölkstockning	56	0.0%	28.6%	0.0%	0.0%
nackspärr	54	16.7%	37.0%	0.0%	0.0%
ont i magen	52	78.8%	21.2%	0.0%	0.0%
torrhosta	52	53.8%	19.2%	9.6%	0.0%
svullna fötter	49	18.4%	51.0%	4.1%	0.0%
trötthet	43	18.6%	48.8%	0.0%	0.0%
klåda	40	35.0%	50.0%	7.5%	0.0%
övervikt	40	0.0%	0.0%	0.0%	47.5%

Table 10.42: Common sign or symptom queries, with proportion of the clicks related to different aspects on diseases.

Query	Search rounds	Facts	Advice	Procedure	Location
influensa	962	77.7%	6.3%	0.0%	0.0%
magkatarr	649	37.1%	27.9%	4.8%	0.0%
halsfluss	558	27.1%	31.9%	0.0%	0.0%
stroke	339	2.7%	8.0%	0.0%	12.7%
öroninflammation	331	32.0%	29.9%	0.0%	0.0%
njursten	307	84.0%	15.3%	0.0%	0.7%
vattkoppor	302	18.5%	27.2%	0.0%	0.0%
svinkoppor	294	18.0%	21.4%	7.5%	0.0%
ibs	283	78.8%	6.7%	0.0%	0.0%
diabetes	270	7.4%	27.8%	0.0%	10.0%
bensår	241	81.3%	18.7%	0.0%	0.0%
körtelfeber	238	41.6%	22.3%	0.0%	0.0%
gallsten	229	72.5%	17.5%	6.1%	0.0%
als	213	22.5%	11.7%	0.0%	0.0%
hepatit c	209	63.2%	5.7%	0.0%	12.4%

Table 10.43: Common disease or syndrome queries, with proportion of the clicks related to different aspects on diseases.

10.5 Properties of a trustworthy portal III

In the case of queries without answers we presented in section 9.5 a number of principles we believe a trustworthy portal should possess, and based on our findings for queries with many answers we will try to summarise our conclusions in a number of additional principles.

In section 10.1 we restricted our interest to so called 6-queries, since these are ones which may result in challenges in mobile settings where the portal providers aim at trying to reduce the size of the presented list of answers and increase the probability for seekers to easily find the ones of interest.

Our analysis revealed that substance features like longer queries, capitalisation and compounds do not seem to result in fewer answers or lower rank of the chosen answers, hence possibly contradict the seekers' expectations on these features to support a portal in its decision procedures. This may also lead to seekers considering the portal to be irrational, and not trustworthy. Thereby we introduce a principle for a trusted portal to capture this problem.

Principle 7: In general, features like longer queries, compounds and capitalisation should result in fewer answers and lower rank of adequate answers.

Another interesting finding is that suggestions of queries by the portals do not result low-rank answers chosen by the seekers, thereby indicating a risk of seekers wondering what they gained by choosing a suggested query.

Principle 8: A query suggested by a portal should result in lower rank of chosen answers than in the case of similar queries posted by a seeker.

Another problem from a seeker perspective is the already identified inconsequent treatment of onomasiological and semasiological properties of queries, and this was captured by Principles 3 and 4 in section 6.7.

Following the studies of queries, we turned our interest to the answers and the interactions as seen by query–answer pairs, and worth emphasising is that some types of answers more often than others are related to higher ranks, for instance, certain types of procedures and symptoms. But at the same time interests in diseases, drugs and anatomical aspects more seldom result in seeker challenges. Among the interests more often originating in 6-queries we find the ones related to specific health care units, which highlighted a question of the use of context information to constrain answer sets. This is also reflected in Principle 1 of a trusted portal.

Moreover, the importance of matching seeker and portal vocabularies, as stressed by Principle 6 in section 9.5, was also seen in the analysis of queries with many answers.

Finally, and possibly slightly disappointing, the core types of queries re-

lated to interests in diseases and symptoms are also the ones most difficult to describe in terms of general principles. However, our studies show that there is a relation between the type of posted queries and chosen answers, but they are found both at a substance and text level. An interesting scenario found among seekers interested in diagnostic procedures is that there is often an interest in both the procedure as such and the providers offering it. Hence, a portal may benefit from combined answers, as in the case of many diseases, covering both aspects. This could lead to improved seeker trust, and possibly even viewing a portal to be insightful.

11 PRINCIPLES OF A TRUSTWORTHY PORTAL

In chapters 9 and 10 we have studied two important challenges, both from the perspective of seekers and portals, with in some cases no answers and in others difficulties finding the interesting one. As often when basing a study on empirical data it may be difficult to identify patterns of interest and conclusions to be drawn, but by a number of principles presented in sections 6.7, 9.5 and 10.5, we tried to highlight some important aspects of, in our opinion, a trustworthy information portal.

In this chapter we will try to further generalise the presented findings and discussions, utilising the framework of search games. Thereby, we also try to show how the conclusions can be applied in other settings where the interaction among seekers and a portal, and the importance of the latter being viewed as trustworthy, are vital.

In chapter 9 we began with a study of the role and impact of location- and time-dependency of interactions, and ended up with two principles (Principles 1 and 5) of a trustworthy (health) information portal considering these aspects. We also saw how certain professions, organisations and activities are location- and time-dependent in the sense that they are only available at certain times and places. Clearly these types of concepts may occur both as queries and answers, and we call them time- and location-dependent to emphasise this property to be accounted for by a (health) information portal to be considered trustworthy be seekers.

Principle I: A (health) information portal's preference relation should take into account time- and location-dependency, but ensure its bounded rationality is not too limited in time and space.

The next challenge dealt with the different types of deviations resulting in portal interpretation problems, which led to two principles (Principle 2 and 6) on the treatment of deviations and vocabularies. As seen in our analysis many of the deviations may be so called second-order ones, i.e. not possible to correct by seekers even if they are told their queries are incorrect. Consequently, a trustworthy portal need to implement a methodology to identify and correct

deviations, in a similar way as humans often are able to correct typographic and other problems as part of their communication. This is related to the vocabularies of the seeker and portal, and since these have been seen to be partly different due to seekers not being familiar with a medical language in general, and especially the subset used by a portal.

Principle II: A (health) information portal's preference relation should be based on a common seeker and portal vocabulary, and able to detect and interpret linguistic second-order deviations.

As in every type of vocabulary, and language, concepts are viewed by seekers as related as, for instance, synonyms and hypernyms. Moreover, in some cases a seeker and portal may differ on the interpretation of a word, due to polysemy, and this may also be a source of distrust in cases when the seeker is not aware of the different meanings. Still, as in the case of human interactions, the portal needs to treat these types of linguistic properties in a sensible way to maintain the seeker's trust. These aspects were covered in Principles 3, 4 and 6, and can be summarised in a general principle for (health) information portals.

Principle III: A (health) information portal's preference relation should respect onomasiological and semasiological relations in the seeker and portal vocabularies.

Finally, we defined two principles (Principles 7 and 8) related to two assumptions on the interactions as such between two actors, where in the first case if the seeker in her opinion "constrains" her query to allow the portal to easier decide its "replies" and if this does not seem to happen she may view the portal as behaving irrationally. The constraints may be explicit, or implicit, in the form of, for instance, longer queries and the use of capital letters, well-defined acronyms, named entities and compounds. The latter principle was related to aspects also noted in the case of queries without answers that when the portal makes suggestions or instructs the seeker on queries, these have to result in the seeker viewing the suggestions as leading to better answers than if not following the advice.

Principle IV: A (health) information portal's preference relation should ensure seeker constraints to result in fewer and/or lower rank of adequate answers.

Principle V: A (health) information portal's preference relation should ensure portal instructions and suggestions to result in fewer and/or lower rank of adequate answers.

However, even with these principles we saw at the end of chapter 10 that in

the case of seekers posting queries on symptoms and diseases their interest varied among facts, advice, procedures etc, and in some cases when topics were gathered by the portal in thematic groups like pregnancy and patient groups these would be yet another area of interest. Another important finding was that the ease with which seekers found the type of answer of interest often shifted between seemingly similar concepts, and one can claim that this just reflects a human behaviour – some queries are easier to answer than others. However, a seeker probably views the portal as an "oracle" in the sense that its vocabulary and knowledge exceed the seeker's and that especially in its core areas, like diseases and symptoms, it should have no problems to provide adequate answers. Hence, it might not be the portal's ability to answer which may impact the seeker's trust, but its ability to not fluctuate in its behaviour. This property is related to the principles above, and we believe that a portal which adheres to these principles will also avoid the outlined problems.

Finally, the last principle is the most important one to try to establish and maintain a seeker's trust in a (health) information portal, and also the basis of search games.

Principle 0: A (health) information portal's preference relation should be in compliance with the intentions of a bounded rationality.

These six principles do, in our opinion, describe important features of a trustworthy (health) information portal, and with these as a guideline in analysis of existing portals and development of improvements or new ones, many of the challenges presented in this thesis in the setting of the portals 1177.se and vardguiden.se would be possible to avoid.

Part III

Summary

12 Discussion and conclusions

We began this thesis with a hypothetical example where Julia searched for information on her symptoms stiff neck and fever at vardguiden.se, and in this concluding chapter we will in the light of our presented efforts return to this example. Moreover, we will elaborate on some open and future challenges not covered in this work.

12.1 Is the future already here?

In chapter 2 we described today's health care and the vision of internet playing an active role in the health care to provide information and advice on diseases and treatments, as well as to support the public's management of their interactions with the health care system. Clearly, a portal will never be able to take the role of a human health care professional, and attempts to introduce internet as the first point of contact may even lead to more visits to health care units. Still the signals are clear, the future care seeker will, and wants to, communicate with a portal on her concerns and problems. However, as discussed this is not a trivial task and at the heart of any such efforts we find the notion of trust with people relying on the information and advice provided by the official health information portals.

As elaborated on in chapter 4, the concept of trust comes in many flavours but they all share the view that it is important, and even vital, in any cooperative activities. Establishing and maintaining trust is difficult among humans, but most probably even more so when one of the actors is a portal and the only way of communication is by queries and answers. Hence, in the setting of (health) information search, trust has to be established and maintained only by the use of a limited language understood by the seeker and portal. To address this challenge we introduced the concept of rationality, capturing the idea that an actor's line of reasoning is rational whenever she is able to choose among any two alternatives which one to prefer, and she will never end up in circular justifications of her choices. Hence, a rational preference relation results, in our

opinion, in a trustworthy behaviour. However, it is important to remember that trust is not about like and dislike, but a way of acting indicating a predictability and if combined with a liking of the behaviour in a willingness to share information and to depend on the other actor. Consequently, a rational preference relation can be seen as one possible way to try to formalise the concept of trust among actors, even when one of them is an internet portal.

Returning to our health information seeker, she posts a query describing her needs and expects the portal to return answers to address these needs, and probably the fewer and more adequate answers provided the more the seeker will view the portal as "competent and rational" and acting in the interest of the seeker. Hence, she trusts the portal, and its potential advice on treatments or information to alleviate her concerns. This view of information search as an interaction of queries and answers between two rational actors, was in chapters 4 and 5 packaged in the conceptual framework of search games, originating in theory of mathematical models of economical reasoning. The aim of the framework was not the formalisation as such, even though it may provide detailed theoretical machinery for analysis, but to provide a platform for descriptive and predictive discussions which are founded in a few basic assumptions and with a terminology feeling familiar – the one of parlour games.

Equipped with the framework of search games, we began an analysis of the official Swedish health information portals 1177.se and vardguiden.se to try to understand how seekers make use of these portals, and how well the portals are able to satisfy the needs of the seekers. Obviously, there is an infinite number of different ways to describe and analyse the interactions of seekers and portals and we decided on two, in our opinion, fundamental areas of interest – queries without answers and ones with many answers. The first topic is fundamental in the context of trust, since a seeker would not post a query if she expected an empty list of answers to be provided by the portal. Hence, any deviations from this may in the worst case result in distrust in the portal's ability to satisfy the seeker's needs and, considering the role of the portals, even potential health risks. Turning to the second topic of interest, many answers to a query is, in general, not a problem per se, but if the seeker has to browse several answers before finding one of interest we may end up in a similar situation of distrust as in the first case. Moreover, by the explosion of the use of mobile devices as smartphones and tablets to search for information, the technology invites to more condensed interactions of higher quality. In part II of this thesis we analysed the search logs of the portals to address these topics of interest, and it provided, to our knowledge, the first description of the use of official and public Swedish health information portals. Moreover, the findings resulted in a number of principles which were summarised in chapter 11 as a guideline for any portal information provider aiming at a trustworthy relation with its

seekers.

Returning to the question in the Prologue if a portal could have done better than in the provided example to support the needs of our fictitious seeker Julia in her efforts to decide whether to seek care, we believe that our studies and discussions show that even with limited efforts, improvements in the decision procedure on answers to present are possible. Ones that will also probably lead to increased trust in the portal and its behaviour. However, our analysis also showed that efforts as the use of semantic resources like medical and other computerised sources to annotate queries and answers to reveal further information to be used to improve the decision procedures is not as straightforward as might be indicated in the literature. The major problem being the vocabularies of these resources to only cover a minor part of the seekers' vocabulary, and in addition in many cases the actually used language of the seekers is full of different types of second-order linguistic deviations to be managed by a portal in a similar way as humans are able to do in their communication.

To conclude, the future is not yet here with an internet portal being able to replace the interaction of an information seeker and a human expert, and still several obstacles remain if this is the path we want to take towards the health care of the future.

12.2 Contributions and reflections

When we embarked on this journey we did not know where it would take us, and even where it would begin. However, as often we started with the raw data, in this case with a number of search logs and the more we studied them, the more peculiarities we found, and the time to clean them increased severalfold. Hence, the first lesson learned is that search logs are captured for certain purposes, but search behaviour analysis may not be one of them and in future efforts this has to be accounted for in a better way – your analysis results will never be better than the quality of the data you begin with. As a consequence, we included chapters 6 and 8 on the material and methods applied in our efforts, but also as a guideline for future analysis.

A second major challenge was if, and how, to place our analysis in a framework to provide a basis for discussions of the material, but also to establish an environment to allow future theoretical as well as practical discussions of our and other similar datasets. Our original intention was not to invent yet another framework, but when studying some of the existing ones they lacked the equal treatment of seekers and portals as well as a model able to attract theoreticians and practitioners both in the fields of computer science, linguistics and health care. It might be viewed as a paradox that our choice was to adapt the frame-

work of game theory with a foundation in the early 20th century's interest in parlour games, mathematical formalism and economics, and by many considered too abstract and difficult to grasp. Still, as discussed in the thesis, game theory has been used in similar settings to study semantics, pragmatics and even induce games from search logs.

The third choice was to re-use the terminology of error analysis and stylistics to describe aspects related to both searches without any answers and ones with many. We could probably have chosen several other points of view and methods, but these turned out to, in our opinion, work well for the purpose of characterising queries and answers, including deviations in a way to open up for further efforts on, for instance, the origin of the different types of deviations and how to manage them. It also included the use of sources like the UMLS to further strengthen the analysis, but as discussed in the previous section, much work remains to be done before we believe sources like this to be useful in a substantial way to establish and maintain trust in information portals.

We may summarise our contributions in this work as providing

- A *model* for descriptive and predictive analysis of (health) *information search*, allowing, in theory, both automatic induction of preference relations and what-if analysis of changed behaviour of seekers and portals
- A *definition* of seeker and portal strategy *preference* facilitating studies of *trust* in (health) information search based on search log data
- An in-depth *analysis* and description of the search carried out at the official Swedish health information portals 1177.se and vardguiden.se, with emphasis on *queries without answers* and *queries with many answers*
- An introduction to error analysis and stylistics for information search analysis
- Examples and discussion on the use of annotation sources as the UMLS for information search analysis and improvements
- Discussion on properties of search logs to facilitate information search analysis
- A set of *principles* a *trustworthy* (health) *information portal*, in our opinion, should adhere to

Since the purpose of the thesis became both to describe the use of two Swedish health portals and to try to provide guidelines on future analysis and on properties recommended to be adhered to by a trustworthy portal, the thesis might, by some readers, be seen as a mesh of threads where some are longer and thicker

and others are shorter and thinner. However, since you as a reader has reached the end we would like to finish with a few questions for future research:

- Is it practically possible to automatically induce a preference relation from a search log, and would it be a rational relation supporting our notion of trust?
- Is it practically possible to define and implement a preference relation satisfying the proposed principles of a trustworthy portal?
- How may sources as the UMLS be better used in a health information setting?
- According to the theory of games, a portal may by its behaviour change the one of seekers, but if and how may this be done? For instance, to achieve fewer queries without answers due to misspellings.
- Rationality, trust and willingness to depend on and take advice from are clearly related, but how can a portal convince a seeker that it is trustworthy for queries regarding fever and coughing when to many seekers it may seem irrational not being able to answer questions on smallpox and borelia (misspelling)?
- Are there any substantial differences in what type of information is searched for and the problems encountered in mobile versus non-mobile search?

REFERENCES

Allott, Nicholas 2006. Game theory and communication. Anton Benz (ed.), *Game theory and pragmatics*, 123–151. Basingstoke: Palgrave Macmillan.

Aumann, Robert J. 2008. *The new Palgrave dictionary of economics*. 2. Steven N. Durlauf and Lawrence E. Blume (eds). Basingstoke: Palgrave Macmillan.

Barros, Gustavo 2010. Herbert A. Simon and the concept of rationality: boundaries and procedures. *Revista de economia política* 30 (3): 455–472.

Barwise, Jon 1989. *The situation in logic*. Stanford, Ca: Center for the Study of Language and Information.

Barwise, Jon and John Perry 1983. *Situations and attitudes*. Cambridge, Mass.: MIT Press.

Bates, Marcia J. 2010. Information behavior. Marcia J. Bates and Mary Niles Maack (eds), *Encyclopedia of library and information sciences*, 2381–2391. Taylor & Francis.

Belkin, N. J., D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan and C. Cool 2003. Query length in interactive information retrieval. *Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval*, SIGIR '03, 205–212. New York, NY, USA: ACM.

Bodoff, David 2009. Emergence of terminological conventions as a searcher-indexer coordination game. *Journal of the American Society for Information Science and Technology* 60 (12): 2509–2529.

Borin, Lars, Markus Forsberg and Lennart Lönngren 2013. SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation* 47 (4): 1191–1211.

Carter, Ronald and Paul Simpson 1989. *Language, discourse and literature: an introductory reader in discourse stylistics*. London: Unwin Hyman.

CeHis, Center för eHälsa i samverkan 2012. *Handlingsplan 2013-2018*. Center för eHälsa i samverkan.

Church, Karen and Nuria Oliver 2011. Understanding mobile web and mobi-

le search use in today's dynamic mobile landscape. *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, MobileHCI '11, 67–76. New York, NY, USA: Association for Computing Machinery.

Crystal, David 1970. New perspectives for language study. 1: stylistics. *English language teaching* 24 (2): 99–106.

Dalianis, Hercules 2002. Evaluating a spelling support in a search engine. Birger Andersson, Maria Bergholtz and Paul Johannesson (eds), *Lecture notes in computer science*, Volume 2553, 183–190. Springer Verlag.

Dalianis, Hercules 2005. Improving search engine retrieval using a compound splitter for Swedish. *Proceedings of the 15th Nordic conference on computational linguistics*, NODALIDA '05.

Devlin, Keith 2006. Situation theory and situation semantics. Dov M. Gabbay and John Woods (eds), *Handbook of the history of logic. vol. 7, logic and the modalities in the twentieth century*, 601–664. Amsterdam: Elsevier North Holland.

Devlin, Keith J. 1991. *Logic and information*. Cambridge: Cambridge University Press.

Digitaliseringskommisionen 2013. *En digital agenda i människans tjänst - sveriges digitala ekosystem, dess aktörer och drivkrafter: delbetänkande*. Stockholm: Fritze.

Dinet, Jerome, Aline Chevalier and Andre Tricot 2012. Information search activity: an overview. *European review of applied psychology : Revue Européenne de psychologie appliquée* 62: 49–62.

Eklund, Ann-Marie 2012a. Tracking changes in search behaviour at a health web site. John Mantas, Stig Kjaer Andersen, Maria Cristina Mazzoleni, Bernd Blobel, Silvana Quaglini and Anne Moen (eds), *Quality of life through quality of information, Proceedings of the 24th European medical informatics conference*, MIE '12, 858–862. IOS Press.

Eklund, Ann-Marie 2012b. Are prepositions and conjunctions necessary in health web searches? *Proceedings of the fourth Swedish language technology conference*, SLTC '12, 23–24.

Eklund, Ann-Marie 2012c. Why query annotations may help in providing accurate public health information. *Proceedings of the fifth workshop on exploiting semantic annotations in information retrieval*, ESAIR'12, 5–6.

Eklund, Ann-Marie 2013a. Mobility and health information searches – a Swedish perspective. Christoph Ulrich Lehmann, Elske Ammenwerth and Christian Nohr (eds), *Proceedings of the 14th world congress on medical and health informatics*, Medinfo '13, 1079–1079. IOS Press.

Eklund, Ann-Marie 2013b. On challenges with mobile e-health – lessons from a game-theoretic perspective. Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, Rajeev Rastogi and Fabrizio Silvestri (eds), *Proceedings of the 22nd ACM international conference on information & knowledge management, CIKM'13*, 1249–1252.

Eklund, Ann-Marie and Dimitrios Kokkinakis 2012. Drug interests revealed by a public health portal. *Proceedings of the Swedish language technology conference workshop: Exploratory query-log analysis.*, SLTC '12, 2.

Eriksson, Håkan and Peter Majanen 2012. *Patient.nu: med vården som hälsoleverantör och internet som vårdcoach*. Lund: Studentlitteratur.

Eysenbach, Gunther 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA annual symposium proceedings*, 244–248. American Medical Informatics Association.

Findahl, Olle 2010. *Svenskarna och internet 2010*. Hudiksvall: World Internet Institute.

Findahl, Olle 2011. *Svenskarna och internet 2011*. Stockholm: .SE (Stiftelsen för internetinfrastruktur).

Findahl, Olle 2012. *Svenskarna och internet 2012*. Stockholm: .SE (Stiftelsen för internetinfrastruktur).

Findahl, Olle 2013. *Svenskarna och internet 2013*. Stockholm: .SE (Stiftelsen för internetinfrastruktur).

Försäkringskassan 2011. *Nya ohälsomått inom sjukförsäkringen*. Försäkringskassan.

Försäkringskassan 2012. Ohälsotalet. www.forsakringskassan.se.

Fox, Susannah 2013. Health fact sheet. www.pewinternet.org.

Fox, Susannah and Maeve Duggan 2013. Health online 2013. Technical Report, Pew Research Center's Internet & American Life Project.

Goatly, Andrew 2008. *Explorations in stylistics*. London: Equinox Publishing.

Hirst, Graeme 2006. Views of text-meaning in computational linguistics: Past, present, and future. Gordana Dodig-Crnkovic and Susan Stuart (eds), *Computing, philosophy, and cognitive science*. Cambridge Scholars Press.

Howard, Nigel 1994a. Drama theory and its relation to game theory. part 1: dramatic resolution vs. rational solution. *Group decision and negotiation* 3: 187–206.

Howard, Nigel 1994b. Drama theory and its relation to game theory. part 2: formal model of the resolution process. *Group decision and negotiation* 3: 207–235.

Hulth, Anette 2013. Webbsök 2013/2014. smittskyddsinstitutet.se.

Hyttsten, Ellen 2012. *Förvaltningsplan för 1177.se verksamhetsåret 2013*. Center för eHälsa i samverkan.

IEP 2012. Internet Encyclopedia of Philosophy. iep.utm.edu.

Ingwersen, Peter and Kalervo Järvelin 2005. *The turn: integration of information seeking and retrieval in context*. Dordrecht: Springer Verlag.

Jain, Shaili, Yiling Chen and David C. Parkes 2009. Designing incentives for online question and answer forums. *Proceedings of the 10th ACM conference on electronic commerce*, EC '09, 129–138. New York, NY, USA: Association for Computing Machinery.

Jain, Shaili and David C. Parkes 2009. The role of game theory in human computation systems. *Proceedings of the ACM SIGKDD workshop on human computation*, HCOMP '09, 58–61. New York, NY, USA: Association for Computing Machinery.

Jakobson, Roman 1960. Concluding statement: linguistics and poetics. Thomas A. Sebeok (ed.), *Style in language*, 350–385. The Technology Press of Massachusetts Institute of Technology John Wiley & Sons, Inc., New York London.

James, Carl 1998. *Errors in language learning and use: exploring error analysis*. Applied linguistics and language studies. London: Longman.

Järvelin, Anni, Richard Berendsen, Gunnar Eriksson, Preben Hansen, Karin Friberg Heppin, Jussi Karlgren, Vivien Petras, Maria Gäde, Mihai Lupu, Florina Piroi, Alba Garcia Seco de Herrera, Stefan Rietberger and 2013. Deliverable 2.4 Use case inventory and final specification of the evaluation tasks. promise-noe.eu.

Josefsson, Ulrika 2011. *Patienters användning av internet: landskap och vägar för 'coping online'*. Lund: Studentlitteratur.

Kairos Future 2012. eTjänster inom hälsa, vård och omsorg.

Keselman, Alla, Allen C. Browne and David R. Kaufman 2008. Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association*, pp. 484–495.

Kokkinakis, Dimitrios and Ann-Marie Eklund 2013. Query logs as a corpus. Andrew Hardie and Robbie Love (eds), *Proceedings of the corpus linguistics 2013*, 329–330. Lancaster University, Lancaster.

Kummervold, Per Egil and Rolf Wynn 2012. Health information accessed on the internet: the development in 5 European countries. *International journal of telemedicine and applications*, vol. 2012.

Leech, Geoffrey N. 2008. *Language in literature: style and foregrounding*. Harlow, England: Pearson Longman.

Leonard, Robert 1995. From parlor games to social science: von Neumann, Morgenstern, and the creation of game theory, 1928-1994. *Journal of economic literature* 33 (2): 730–761.

Leonard, Robert 2010. *von Neumann, Morgenstern, and the creation of game theory: from chess to social science, 1900–1960*. New York: Cambridge University Press.

Leshem, Shosh and Vernon Trafford 2007. Overlooking the conceptual framework. *Innovations in education and teaching international* 44 (1): 93–105.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady* 10: 707.

Leyton-Brown, Kevin and Yoav Shoham 2008. *Essentials of game theory : a concise, multidisciplinary introduction*. Morgan & Claypool Publishers.

LIF, Läkemedelsindustriföreningens Service AB 2013. Presentation av LIF. fass.se.

Lindskog, Bengt I., Åke Andrén-Sandberg, Urban Frank and Poul Buckhöj 2008. *Medicinsk terminologi*. 5 ed. Stockholm: Norstedts Akademiska.

Liu, Jian, Yiqun Liu, Min Zhang and Shaoping Ma 2013. How do users grow up along with search engines?: a study of long-term users' behavior. Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, Rajeev Rastogi and Fabrizio Silvestri (eds), *Proceedings of the 22nd ACM international conference on information & knowledge management, CIKM'13*, 1795–1800.

Mannberg, Anna 2013. Om 1177 Vårdguiden. 1177.se.

Marshall, Alfred 1920. *Principles of economics. An introductory volume*. 8. ed. London: Macmillan.

McKnight, D. Harrison and Norman L. Chervany 2001. Trust and distrust definitions: one bite at a time. *Trust in cyber-societies*, 27–54. Springer Verlag.

Medlock, Stephanie, Saeid Eslami, Marjan Askari, Danielle Sent, Sophia E. de Rooij and Ameen Abu-Hanna 2013. The consequences of seniors seeking health information using the internet and other sources. *Studies in health technology and informatics* 192: 457–460.

Miles, Matthew B., A. Michael Huberman and Johnny Saldana 2014. *Qualitative data analysis: a methods sourcebook*. 3. ed. Los Angeles: SAGE Publications.

Moradi, Farnaz, Ann-Marie Eklund, Dimitrios Kokkinakis, Tomas Olovsson and Philippas Tsigas 2014. A graph-based analysis of medical queries of a Swedish health care portal. *Proceedings of the 5th international workshop on health text mining and information analysis*, Louhi '14, 2–10. Gothenburg, Sweden: Association for Computational Linguistics.

Myerson, Roger B. 1991. *Game theory: analysis of conflict*. Cambridge, Mass.: Harvard University Press.

von Neumann, John 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.

von Neumann, John and Oskar Morgenstern 1944. *Theory of games and economic behavior.* 60th anniversary ed. Princeton, NJ: Princeton University Press.

NLM, National Library of Medicine 2013a. Unified Medical Language System, UMLS. nlm.nih.gov.

NLM, National Library of Medicine 2013b. Fact sheet Medical Subject Headings. nlm.nih.gov.

NLM, National Library of Medicine 2014. Unified Medical Language System. nlm.nih.gov.

Oelke, Daniela, Ann-Marie Eklund, Svetoslav Marinov and Dimitrios Kokkinakis 2012. Visual analytics and the language of web query logs - a terminology perspective. *The 15th EURALEX International Congress (European Association of Lexicography)*, EURALEX '12, 541–548.

Palen, Ted E., Colleen Ross, J. David Powers and Stanley Xu 2012. Association of online patient access to clinicians and medical records with use of clinical services. *Journal of the American Medical Association* 308 (19): 2012–2019.

Parfionov, George and Roman R. Zapatrin 2011. Memento ludi: information retrieval from a game-theoretic perspective. Leon A. Petrosyan and Nikolay A. Zenkevich (eds), *Contributions to game theory and management volume IV*, 339–246. Graduate school of management St. Petersburg University.

Parikh, Prashant 2010. *Language and equilibrium*. Cambridge, Mass.: The MIT Press.

Rahmqvist, Mikael and Ana-Claudia Bara 2007. Patients retrieving additional information via the internet: a trend analysis in a Swedish population, 2000-05. *Scandinavian journal of public health* 35 (5): 533–539.

Reis, Sofia, Karen Church and Nuria Oliver 2012. Rethinking mobile search: towards casual, shared, social mobile search experiences. *Proceedings of the Searching 4 Fun! workshop*.

Reiter, Raymond 1978. On closed world data bases. Herve Gallaire and Jack Minker (eds), *Logic and data bases*, 119–140. New York: Plenum Press.

Robson, Colin 2011. *Real world research: a resource for users of social research methods in applied settings*. 3. ed. Chichester: Wiley.

Rudestam, Kjell Erik and Rae R. Newton 2007. *Surviving your dissertation: a comprehensive guide to content and process*. 3. ed. Los Angeles: SAGE Publications.

Sadasivam, Rajani S., Rebecca L. Kinney, Stephenie C. Lemon, Stephanie L. Shimada, Jeroan J. Allison and Thomas K. Houston 2012. Internet health information seeking is a team sport: Analysis of the Pew Internet Survey. *International journal of medical informatics* 82: 193–200.

de Saussure, Louis 2007. Pragmatic issues in discourse analysis. *Critical approaches to discourse analysis across disciplines* 1: 179–195.

SCB 2014. *Statistisk årsbok för sverige*. Statistiska centralbyrån.

Simon, Herbert A 1997. *An empirically based microeconomics*. Cambridge: Cambridge University Press.

Skeppstedt, Maria, Maria Kvist and Hercules Dalianis 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. *Proceedings of the eighth international conference on language resources and evaluation*, LREC '12, 1250–1257. Istanbul.

Slantchev, Branislav L. 2007. Game theory: preferences and expected utility. Technical Report, Political Science Courses, University of California–San Diego.

SLL 2010. Över 100 000 besök på Vårdguiden i mobilen. Stockholms läns landsting.

SLL 2013. Nu startar Stockholms läns landstings nya e-hälsouppdrag i samarbete med alla landsting och regioner. Stockholms läns landsting.

Socialdepartementet 2010. Nationell eHälsa – strategin för tillgänglig och säker information inom vård och omsorg. regeringen.se, Stockholm.

Socialstyrelsen 2013. Om Snomed CT. socialstyrelsen.se.

Stockholms stad 2013. *Statistisk årsbok för Stockholm 2014*. Stockholm: Utrednings- och statistikkontoret, Stockholms stad.

sundhed.dk 2011. Background and status report. sundhed.dk.

Svenska Akademien 2006. *Svenska Akademiens ordlista över svenska språket*. 13 ed. Stockholm: Svenska Akademien.

The Guardian 2013. Google introduces the biggest algorithm change in three years. 27 Sep 2013.

Tucker, Albert W., R. Duncan Luce and Harry Waldo Kuhn 1959. *Contributions to the theory of games. 4*. Princeton, N.J.: Princeton University Press.

Wang, Peiling 2011. Information behavior and seeking. Ian Ruthven and Diane Kelly (eds), *Interactive information seeking, behaviour and retrieval*, 15–41. Facet publishing.

Wangberg, Silje, Hege Andreassen, Per Kummervold, Rolf Wynn and Tove Sorensen 2009. Use of the internet for health purposes: trends in Norway 2000–2010. *Scandinavian journal of caring sciences* 23 (4): 691–696.

Weaver, James B., Darren Mays, Gregg Lindner, Dogan Eroglu, Frederick Fridinger and Jay M. Bernhardt 2009. Profiling characteristics of internet medical information users. *Journal of the American Medical Informatics Association* 16 (5): 714–722.

Weaver, James B., Darren Mays, Stephanie Sargent Weaver, Gary L. Hopkins, Dogan Eroglu and Jay M. Bernhardt 2010. Health information-seeking behaviors, health indicators, and health risks. *American journal of public health : official journal of the American Public Health Association* 100 (8): 1520–1525.

Wikipedia 2013. wikipedia.org.

Zermelo, Ernest 1912. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels (On the application of set theory to the theory of chessgames). *Proceedings of the fifth international congress of mathematicians*, 501–504.

INDEX

A SUPPLEMENTARY MATERIAL

This appendix contains additional information referenced in the thesis.

Query	Utterances
akutmottagning	76,550
gynekolog	47,238
vårdcentral	31,195
psykiatrisk öppenvårdsmottagning	24,150
ögon	22,696
vattkoppor	22,394
barnmorskemottagning	21,655
feber	20,330
sex och samlevnad	19,327
barnavårdscentral	19,038
närakut	18,496
säsongsinfluensan vaccination	18,414
halsfluss	17,225
influensa	16,885
diabetes	16,156
urinvägsinfektion	15,735
gravid	14,880
sjukgymnast	14,618
utslag	14,475
hosta	14,371

Table A.1: Most common queries Stockholm.

Query	**Utterances**
vårdcentral	15,700
barnmorskemottagning	8,920
feber	5,999
gynekolog	5,679
gravid	5,451
abortmottagning	5,294
förlossning	5,091
influensa	4,021
gynekologimottagning	3,828
akutmottagning	3,553
närakut	3,531
halsfluss	3,499
vattkoppor	3,389
hosta	3,363
akutmottagning sjukhus vuxen	3,142
diarre	3,107
magsjuka	2,908
sex och samlevnadsmottagning	2,795
barnavårdscentral	2,780
urinvägsinfektion	2,724

Table A.2: Most common queries Stockholm Mobile.

Query	**Utterances**
vårdcentral	498,856
tandvård	224,502
cancer	129,997
barnavårdscentral	99,777
barnmorskemottagning	97,847
hälsocentral	85,686
sjukhus	82,279
akutmottagning	72,178
psykiatrisk vård	55,056
ungdomsmottagning	45,590
mödravård	35,173
jourcentral	29,902
närakuten	28,341
sex och samlevnad	28,319
familjeläkarmottagning	25,496
sjukgymnast	25,225
psykiatri	24,974
sjukgymnastik	22,786
diabetes	20,340
distriktssköterskemottagning	19,572

Table A.3: Most common queries Sweden.

Query	Utterances
mina vårdkontakter	359
egenremiss	246
sjukresor	220
vaccination	211
gravid	182
urinvägsinfektion	181
feber	177
influensa	174
vårdgaranti	171
byta vårdcentral	167
webbisar	154
halsfluss	148
mammografi	146
diabetes	132
vårdval	125
klamydia	121
stroke	118
blanketter	99
hosta	96
vattkoppor	96

Table A.4: Most common queries Västra Götaland.

	Stockholm	Sweden	Västra Götaland
Time period	2010-10 – 2011-09	2011-09 – 2013-09	2010-06 – 2011-09
Utterances	6,539,095	8,600,596	31,054
Queries	874,283	967,219	8,818
Sessions	–	–	27,881

Table A.5: Basic search log statistics.

Age	**Young (0–19)**	**Adult (20–64)**	**Elderly (65+)**
Max	30.1	70.4	20.0
Hägersten-Liljeholmen	20.3	67.4	12.2
Skarpnäck	22.8	65.9	11.3
Östermalm	15.9	64.1	20.0
Average	22.4	63.4	14.0
Rinkeby-Kista	27.3	62.7	10.0
Min	14.8	58.3	10.0

Table A.6: Age distribution for Stockholm.

Age	**Young (0–24)**	**Adult (25–64)**	**Elderly (65+)**
Max	31.5	54.1	23.5
Stockholm	30.5	54.1	15.4
Västra Götaland	29.9	51.5	28.6
Average	29.3	49.9	20.9
Östergötland	30.4	49.9	19.7
Jämtland	28.3	49.6	22.1
Min	27.8	48.7	15.4

Table A.7: Age distribution for Sweden.

Education	**< College**	**College**	**> College**
Max	27.2	41.8	65.2
Östermalm	7.2	24.4	65.2
Hägersten-Liljeholmen	11.9	34.1	52.2
Average	15.9	34.0	46.8
Skarpnäck	14.9	34.9	47.7
Rinkeby-Kista	27.2	35.0	30.2
Min	7.2	24.4	29.6

Table A.8: Education distribution for Stockholm.

Education	< College	College	> College
Max	27.8	50.5	40.6
Stockholm	18.5	38.1	40.6
Jämtland	23.3	48.9	26.4
Östergötland	24.4	43.8	30.4
Average	24.4	45.4	28.4
Västra Götaland	24.0	42.7	31.5
Min	18.5	38.1	23.5

Table A.9: Education distribution for Sweden.

Immigration	Percentage
Max	56.6
Rinkeby-Kista	56.6
Average	25.2
Skarpnäck	20.8
Östermalm	17.3
Hägersten-Liljeholm	17.0
Min	14.1

Table A.10: Immigrant distribution for Stockholm.

Immigration	Percentage
Max	22.2
Stockholm	22.2
Västra Götaland	15.3
Östergötland	12.6
Average	12.1
Jämtland	6.9
Min	5.0

Table A.11: Immigrant distribution for Sweden.

Age	**16–29**	**30–44**	**45–49**	**50–54**	**55–59**	**60–64**
Max	8.1	15.5	40.6	66.1	93.3	109.7
Rinkeby-Kista	6.5	15.5	40.6	66.1	93.3	109.7
Skarpnäck	7.0	11.0	25.5	37.1	47.3	80.5
Average	5.9	10.3	23.3	35.9	50.6	68.1
Hägersten-Liljeholmen	5.1	9.1	21.6	33.8	46.3	62.7
Östermalm	3.0	6.9	15.1	21.3	27.9	40.9
Min	3.0	5.6	14.5	21.3	27.9	40.9

Table A.12: Ohälsotal for Stockholm.

Age	**16–29**	**30–49**	**50–59**	**60–64**
Max	19.3	46.9	51.7	83.8
Jämtland	16.6	45.5	51.7	82.7
Västra Götaland	13.6	42.6	51.3	77.4
Average	14.5	41.2	47.1	73.0
Östergötland	13.1	39.1	47.8	73.8
Stockholm	10.8	30.1	41.8	66.5
Min	10.8	30.1	40.4	61.2

Table A.13: Ohälsotal for Sweden.