What customers really want to know from tourism information systems but never dared to ask

Michael Dittenbach[*], Dieter Merkl[*†], and Helmut Berger[*]

[*] E-Commerce Competence Center – EC3
Donau-City-Straße 1, A–1220 Wien, Austria
{michael.dittenbach,helmut.berger}@ec3.at

[†] Technische Universität Wien
Institut für Softwaretechnik und Interaktive Systeme
Favoritenstraße 9–11/188, A–1040 Wien, Austria
dieter@ifs.tuwien.ac.at

Abstract

We have developed a natural language query interface for the largest Austrian web-based tourism platform, *Tiscover*. During a period of 10 days this interface was made publicly available by means of an advertisement on and a hyperlink from *Tiscover*'s homepage. In this paper, we will describe the results and our insights from analysing the natural language queries collected during this field trial. This analysis shows how users formulate queries when their imagination is not limited by a conventional search interfaces with structured forms consisting of check boxes, radio buttons and special-purpose text fields. The results of this field test are thus valuable indicators into which direction the web-based tourism information system should be extended to better serve the customers.

Keywords Tourism information system, Natural language processing, User behavior study, User interface, Web-based information system

1 Introduction

The development and availability of efficient and appropriate search functions are still a challenge in the field of database and information systems. Consider, for example, the context of tourism information systems where intuitive search functionality plays a crucial role for the economic success. Querying an information system in natural language is especially appealing in the tourism domain because users usually have very different backgrounds regarding computer literacy. Hardly any computer scientist or technically interested person has problems understanding the Boolean logic underlying conventional web search engines. Unfortunately, a growing majority of people using search engines has.

An analysis of query logs of the search engine *Excite* has shown that, in practice, only 9% of the queries contain Boolean operators or the modifiers "+" and "−" [5]. The latter two require that a query term must or must not be present in the searched pages. Although large web search engines like *Google*, *Altavista* and of course thousands of smaller site-specific search facilities have the same superficial appearance, they tend to interpret queries with subtle differences that can lead to searches not meeting the user's intention. Without further information, one cannot be sure if a query is treated case sensitive or not, or how the keywords are connected logically, i.e. if all or any of the terms have to apply [12].

To take away the fear of this rather technical way of searching for information, natural language should present a convenient form of interaction with such systems. In particular, we foresee the following benefits for the user. She or he is relieved from the burden of having to learn and to use either strictly logical or highly structured query languages. The user could interact naturally with the system, using her or his style of description of the needed information. Obviously, this

should be expected to be of special importance in the tourism sector where people are often characterized by having rather unstructured imagination of their information need [10].

Hence, we have developed a natural language interface for the largest Austrian web-based tourism platform *Tiscover* (*http://www.tiscover.com*) [11]. *Tiscover* is a well-known tourism information system and booking service in Europe that already covers more than 50,000 accommodations in Austria, Germany, Liechtenstein and Switzerland. It integrates a variety of additional services like live weather reports, event booking, special holiday package offers, route planning and a job market.

More specifically, our natural language interface allows users to search for accommodations throughout Austria by formulating the query in a natural language sentence either in German or English [2]. The language of the query is automatically detected and the result is presented in the respective language. For the task of natural language query analysis we followed the assumption that shallow natural language processing is sufficient in restricted and well-defined domains [8]. In particular, our approach relies on the selection of query concepts followed by syntactic and semantic analysis of the portion of the natural language query where the concepts appear.

During 10 days of March 2002, we tested the assumptions behind the natural language interface in a field trial where the interface was accessible via a hyperlink from the *Tiscover* homepage. The time for the trial was chosen deliberately because close to vacation periods, as the Easter week in our case, the traffic at a web-based tourism information system is higher than during other times. The major objectives for the field trial were, first, to verify whether or not users accept natural language interaction. That means, we are interested if the users actually type natural language sentences to describe their information needs. Second, we hoped for a broad spectrum of natural language requests for tourism information, now that the users are no longer biased by available tick-boxes, radio buttons or selection lists. Finally, we were interested in the practical performance of the natural language interface given a real-world setting.

In this paper, we will put main emphasis on our findings from the analysis of the natural language queries collected and processed during the field trial. The remainder of this paper is organized as follows. In Section 2 we provide a very brief overview of the steps in natural language query processing. Section 3 outlines the design goals for the interface. Then, Section 4 provides a description of the data collected during the field trial as well as the consequences which can be drawn from this data. Finally, Section 5 gives some conclusions.

2 Natural Language Processing

For the sake of brevity, we will describe in this Section only the logical sequence of processing steps during natural language analysis for our query interface. We refer the reader who is interested in more general aspects of natural language information systems to [1, 4, 6, 7]. For a more thorough technical description of our query interface, we refer to [2].

When a query is sent to the natural language processing module, in a first step, the language of the query has to be identified. Currently we support German and English, but the system has been designed to allow for easy integration of additional languages. The language identification is based on a text classification approach using *n-gram* statistics [3]. The numbers of *n-gram*s in a query are compared with *n-gram* distributions of German and English texts. Depending on the similarity between those, the language is chosen. However, if both languages are nearly equally probable, the system returns that the language of the query can not be determined and asks the user to rephrase her or his query.

Next, to improve the retrieval performance, potential orthographic errors and misspellings have to be considered. Therefore, we used a phonetic algorithm to find and correct such errors, e.g. *"Insbruk"* will be replaced by the correct city's name *"Innsbruck"*.

An important issue regarding tourism information is to automatically identify proper names consisting of more than one word, e.g. *"St. Anton am Arlberg"*, without having the user to enclose it with quotes. This also applies to phrases and multi-word denominations like *"swimming pool"* or *"car park"* to name but a few. We chose a regular expression approach to identify such cases.

In the next query processing step, the relevant concepts and modifiers have to be tagged. For this purpose, we have developed an XML-based ontology covering the semantics of domain specific concepts and modifiers and describing linguistic concepts like synonymy. Additionally, a lightweight grammar describes how certain concepts may be modified by prepositions and adverbial or adjectival structures that are also specified in the ontology.

Finally, the query has to be transformed into an SQL statement to retrieve information from the database. Using the tagged concepts and modifiers together with the rule set and parameterized SQL fragments also defined in the knowledge base allows the construction of a complete SQL statement reflecting the natural language query.

As an example consider the query *"I am looking for a hotl in St. Abton am Arlberg with sauna and a swiming pool. The hotel should furthermore be suitable for children and pets should be allowed."* As can be seen, the query contains several misspellings such as *"hotl"*, *"Abton"* and *"swiming pool"*. After correcting these, the relevant concepts of this sample query are *"hotel"*, *"St. Anton am Arlberg"*, *"sauna"*, *"swimming pool"*, *"suitable for children"* and *"pets allowed"*.

A generic XML description of the matching accommodations is created to allow for device-dependent output, customized according to features like screen size or bandwidth.

Our information system covers a part of the *Tiscover* database, which, as of October 2001, provides access to information about 13,117 Austrian accommodations. These are described by a large number of properties including the respective numbers of various room types, different facilities and services provided in the accommodation, or even the type of food.

These accommodations are located in 1,923 towns and cities that are again described by various features, mainly information about possible sports activities, e.g. mountain biking or skiing, but also the number of inhabitants or the sea level. The federal states of Austria are the higher-level geographical units. For a part of the data, we integrated the geographical coordinates of the cities and towns to additionally provide information about the distance between places. Hence, the system can be queried for accommodations *close* to a certain place.

3 Design Considerations for the Web-Based Interface

Our major design goal at the outset of the project was to provide a simple and easy to use interface. Hence, the interface is dominated by a text-box where the user can enter her or his query and a submit button, the latter one labeled with *"ask"*. During the field trial we additionally provided short textual descriptions in both German and English in form of sample queries. The sample query *"I am looking for a double room in the center of Salzburg with indoor pool."* is the only hint on the capabilities of the interface. The intention was to collect a broad range of accommodation requests and, thus, to find out what the users really want. Our aim was not to bias the users's imagination when formulating a query. This, admittedly, with the risk of disappointing the user when no or just inappropriate results were found. See Figure 1 for a screen-shot of the interface as it looked like during the field trial. For the curious reader we shall note that the interface is still accessible at *http://www.tiscover.at/powersearch*.

Figure 2 shows the conventional interface of *Tiscover* for searching accommodations. The area (federal state, region, city) can be chosen either by typing the name directly into the text field or via clicking through the hierarchy of names of geographical locations. Further criteria are the name of the accommodation, the chain it belongs to and, perhaps, a particular "theme", e.g. family hotel, as well as several amenities the accommodation should provide. Note, this list of amenities is rather small compared to the complete information of the *Tiscover* database to keep the interface concise.

We also implemented the look and feel of the *Tiscover* design in order to avoid distraction from the user's task. On the result screen (see Figure 3), we present the original query as well as the concepts identified by the natural language processing to provide the user with feedback regarding the quality of natural language analysis. Below the list of accommodations matching the criteria, we have provided a feedback form where users can enter a comment and rate the quality of the result. After the field test, it turned out that only 3.37% of the queries have either been annotated or rated where the number of positive and negative comments were nearly equal. Due to the unsupervised nature of the test without any reward for the users, this figure is not surprising because of the additional time it takes to assess the quality of the result and then comment on it. At the bottom of the page, the input field prefilled with the posed query is presented to allow for convenient query reformulation or refinement. About 10% of the queries were modified by adding or deleting parts of the original query.

Figure 1: Natural language query interface

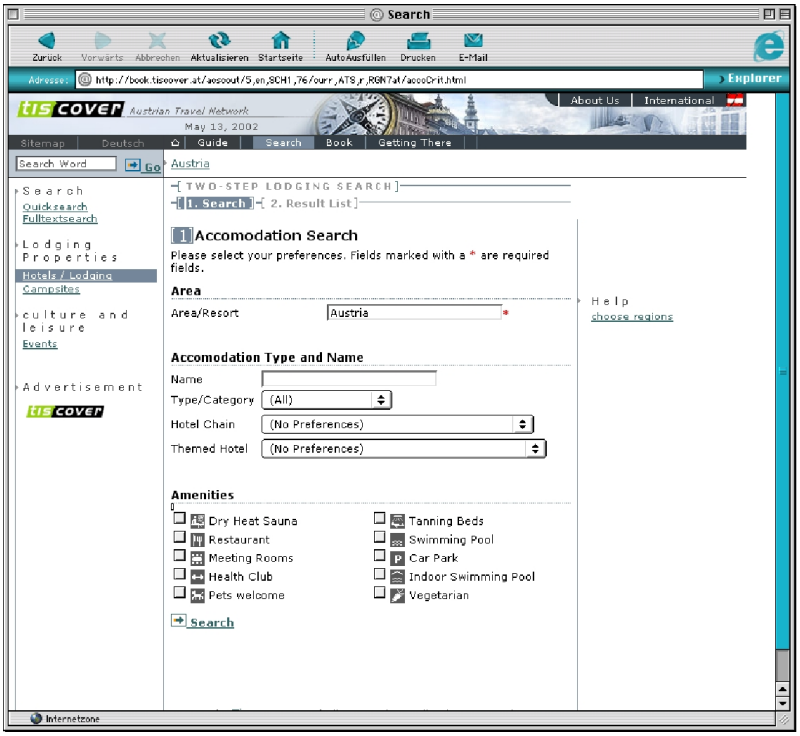

Figure 2: Standard *Tiscover* search interface

Figure 3: Result page with matching accommodations and feedback form

4 Results from the Field Trial

The field test was carried out from March 15 to March 25, 2002. During this time our natural language interface was promoted on and linked from the main *Tiscover* page. We obtained 1,425 unique queries through our interface, i.e. equal queries from the same client host have been reduced to one entry in the query log to eliminate a possible bias for our evaluation of the query complexity.

In Table 1, a list of countries and the respective numbers of queries is shown. Naturally, most of the queries (39.73%) came from Austrian hosts, followed by hosts from the *.net* top-level domain, most of which have been identified as German internet service providers by manual inspection. After the 13.13% of queries from the US commercial domain several European countries can be found. A country could not be assigned to 20.42% of the queries because of a non-resolvable domain name.

Of those 1,425 unique queries, 1,213 (85.12%) were identified as German, 120 (8.42%) were identified as English and 92 (6.46%) were not identifiable, e.g. non-sentence queries like *"hotel salzburg"* that are possible in both languages or just nonsense like *"ghsdfkjg"*. Based on the 1,333 identified queries we found 85 queries that were not in the scope of our natural language interface. Among these were, for example, questions about car rentals and, of course, sex. Obviously, in any kind of publicly available service like this, not all of the people are using it for the intended purpose. However, this number is rather low assuming the rather short description we displayed on the start page to give an idea what kind of information can be queried.

To assess the overall quality of the language identification we manually inspected the submitted natural language queries. For each query the system assigns either the most probable language or considers the language of the query as

# of queries (%)	country	# of queries (%)	country
566 (39.73%)	Austria	6 (0.42%)	Luxembourg
229 (16.07%)	.net (mostly German ISPs)	5 (0.35%)	Hungary
187 (13.13%)	US commercial	4 (0.28%)	Belgium
70 (4.91%)	Germany	2 (0.14%)	South Africa
22 (1.54%)	Switzerland	2 (0.14%)	Australia
17 (1.19%)	Italy	1 (0.07%)	US military
14 (0.98%)	Netherlands	1 (0.07%)	France
8 (0.56%)	UK	291 (20.42%)	unknown (not resolved)

Table 1: Origin of queries (derived from the top-level domain of the accessing host)

being ambiguous. For example, a German query can either be identified correctly, as English or as ambiguous. In Table 2, we provide the actual figures resulting from the manual inspection. Thus, of the 1,213 queries identified as German, 1,210 were correctly identified. However, three in fact English queries have been misclassified as being German. In the third row of the table, we can see that of the 92 queries identified as ambiguous, 74 were actually German. This classification error can be explained by the peculiarities of the language identification algorithm based on *n-grams*. Especially short queries lead to *n-gram* distributions that do not allow to distinguish between English and German with the required accuracy. In total, of the submitted queries 1,306 were in fact German of which 1,210 were correctly identified. This yields an identification accuracy of 92.6% for the German language. The respective result for the English language is 95.1%.

	manual analysis			
	german	**english**	**ambiguous**	
german	1,210	3	0	1,213
english	22	98	0	120
ambiguous	74	2	16	92
totals	**1,306**	**103**		
identification accuracy	**92.6%**	**95.1%**		

Table 2: Manual analysis of language identification accuracy

To provide some technical information, for the 1,333 processed queries, the mean processing time was 2.63 seconds with a standard deviation of 1.42 seconds. The median of 2.27 seconds shows that there were only a few outliers with longer processing times. Given these figures, we can safely say that our system is usable regarding its response time. Even with adding a few seconds for data transmission time over the Internet, the response time still lies below the magic number of ten seconds as suggested by [9]. These ten seconds have been measured in usability studies as the approximate maximum attention span of users when waiting for a web page to be loaded before canceling the request.

We will compare the results of two studies analyzing query log files of the large and popular search engines *Altavista* and *Excite* with the results of our analysis, since only few research papers dealing with user behavior in web searches exist. In [5] and [13] the authors have shown that the average number of words per query is very small, namely 2.35, interestingly the same in both studies. This indicates that most of the people searching for information on the Internet could improve the quality of the results by specifying more query terms. Our field test revealed the encouraging result of an average query length of 8.90 words for German queries, and of 6.53 for the English queries, see Table 3 for details.

In more than a half (57.05%) of the 1,425 queries, users formulated complete, grammatically correct sentences whereas only 21.69% used our interface like a keyword-based search engine. The remaining set of queries (21.26%) were partial sentences like *"double room for 2 nights in Vienna"*. Several of the queries consisted of more than one natural language sentence, e.g. *"We look for a house at one of the lakes in Austria from July 22 until July 28, 2002. We are a familiy with 2 children of 8 and 11 years and have a dog. We are searching for a house with lake entrance."* This approves our assumption that users accept the natural language interface and are willing to type more than just a few keywords to search for information. More than this, a substantial portion of the users is typing complete sentences to express their information needs. Furthermore, the average number of relevant concepts occurring in the German queries is 3.41 with a standard

# of words per query	# of queries (%)	# of words per query	# of queries (%)
1	76 (5.33%)	18	12 (0.84%)
2	92 (6.46%)	19	14 (0.98%)
3	117 (8.21%)	20	6 (0.42%)
4	82 (5.74%)	21	9 (0.63%)
5	109 (7.65%)	22	5 (0.35%)
6	147 (10.32%)	23	8 (0.56%)
7	87 (6.11%)	24	2 (0.14%)
8	105 (7.37%)	25	7 (0.49%)
9	98 (6.88%)	26	2 (0.14%)
10	101 (7.08%)	27	3 (0.21%)
11	66 (4.63%)	28	3 (0.21%)
12	75 (5.25%)	29	2 (0.14%)
13	55 (3.86%)	32	1 (0.07%)
14	53 (3.90%)	35	1 (0.07%)
15	30 (2.11%)	37	3 (0.21%)
16	22 (1.54%)	66	1 (0.07%)
17	28 (1.96%)	76	1 (0.07%)

Table 3: Word occurrence statistic

deviation of 1.96, which is still one word per query more than found in the surveys mentioned above. It can be assumed, that, by formulating a query in natural language, users are more specific than compared to keyword-based searches.

To inspect the complexity of the queries, we considered the number of concepts and the usage of modifiers like *"and"*, *"or"*, *"not"*, *"near"* and some combinations of those as quantitative measures. Table 4 shows the distribution of the numbers of concepts per query. For example, consider row four of this Table. The entries in this row show the number of queries with three concepts. In particular, we have 310 German and 28 English queries. Note that these figures were derived by manual inspection of the users' original natural language queries. The majority of German queries contains one to five concepts relevant to the tourism domain with a few outliers of more than 10 concepts. The latter can be explained by people asking for an accommodation in a specific region by enumerating potentially interesting cities and villages.

	query language		
concepts	**german**	**english**	**totals**
0	47	5	**52**
1	77	28	**105**
2	272	38	**310**
3	310	28	**338**
4	245	12	**257**
5	137	5	**142**
6	49	2	**51**
7	38	1	**39**
8	18	1	**19**
9	11	0	**11**
10	4	0	**4**
11	1	0	**1**
17	3	0	**3**
21	1	0	**1**
totals	**1,213**	**120**	**1,333**

Table 4: Number of concepts per query (counted by manual inspection)

In analogy to Table 4, the Tables 5 (a) and 5 (b) give an indication regarding the quality of the natural language query analysis. In particular, Table 5 (a) provides the numbers of identified concepts per query, whereas Table 5 (b) that of not

	query language		
concepts	**german**	**english**	**totals**
0	71	14	**85**
1	104	27	**131**
2	326	39	**365**
3	312	24	**336**
4	201	10	**211**
5	106	2	**108**
6	50	2	**52**
7	19	2	**21**
8	13	0	**13**
9	6	0	**6**
10	1	0	**1**
16	3	0	**3**
20	1	0	**1**
totals	**1,213**	**120**	**1,333**

(a) Concepts identified by the natural language processing

	query language		
concepts	**german**	**english**	**totals**
0	817	88	**905**
1	348	29	**377**
2	45	3	**48**
3	3	0	**3**
totals	**1,213**	**120**	**1,333**

(b) Concepts not identified by the natural language processing

Table 5: Concepts that have been identified or not identified by the natural language processing module of our interface

identified concepts. Again, the figures given in Table 5 (b) were derived by manual inspection. We shall note that most of the concepts not identified, originated from queries falling into the categories of region names, pricing information, room availability and arrival and departure dates. These informations were not contained in the part of the database used for our natural language system.

Another aspect of the complexity of natural language queries are words connecting concepts logically or modifying their meaning. These modifiers can be compared to operators like "AND", "OR", "+" or "−" of web search engines. In Table 6 (a) we can see that the distribution of occurrences of the modifier *"and"* corresponds to the number of concepts. In 320 queries the modifier *"and"* was used twice which relates to the occurrence of three concepts per query (cf. Table 4). The occurrence statistic includes all implicitly used modifiers *"and"*, i.e. those *and*s that are included because of the resulting SQL statement, as well as those explicitly defined, i.e. those *and*s that are provided with the natural language query. The query *"I am looking for a hotel with sauna, solarium and whirlpool in Tyrol"* includes one explicitly used *"and"*, and three implicit *"and"* modifiers.

Due to the assumption that the underlying semantics of combining concepts is based on the intention to provide facilities somebody wants to have, we defined the *"and"* modifier to be the default logic for combining concepts if no explicitly defined modifier is present. This assumption is made to provide a convenient technique to map the concepts used in a query onto the underlying program logic.

The modifier *"or"* is used far less than *"and"*, as shown in Table 6 (b). In particular, *"or"* is used in 103 queries only. *"Or"* is mostly used to provide a set of locations or types of accommodations of interest, e.g. *"I am looking for a farm or an apartment in Tyrol or Salzburg"*.

An interesting fact is, that the *"not"*-modifier is used in a very small subset of queries (cf. Table 6 (c)). The modifier *"not"* occurs in only 19 German and 3 English queries. This implies, that the vast majority of users formulate their intentions without the need of excluding concepts. In most of the cases where a *"not"* is used to exclude a specific property of a region or an accommodation, users wanted to avoid places where pets are allowed as well as accommodations that are *not* particularly well-suited for children, the latter, perhaps, to stress the desire to find a quiet place. Another common use of *"not"* is to exclude one or more cities from a query where an accommodation in a federal state or region was wanted, e.g. *"I am looking for a hotel in Tyrol, but not in Innbruck and not in Zillertal."*

Table 6 (d) shows the number of occurrences of the modifier *"near"* which has been expressed by terms like *"around"*,

	query language		
and	**german**	**english**	**totals**
1	281	38	**319**
2	320	29	**349**
3	246	11	**257**
4	140	6	**146**
5	41	1	**42**
6	33	1	**34**
7	16	0	**16**
8	4	0	**4**
9	2	0	**2**
10	1	0	**1**
totals	**1,084**	**86**	**1,170**

(a) Usage of modifier *and*

	query language		
or	**german**	**english**	**totals**
1	67	4	**71**
2	18	1	**19**
3	6	1	**7**
6	1	0	**1**
8	1	0	**1**
12	3	0	**3**
16	1	0	**1**
totals	**97**	**6**	**103**

(b) Usage of modifier *or*

	query language		
not	**german**	**english**	**totals**
1	12	3	**15**
2	7	0	**7**
totals	**19**	**3**	**22**

(c) Usage of modifier *not*

	query language		
near	**german**	**english**	**totals**
1	112	9	**121**
2	0	1	**1**
totals	**112**	**10**	**122**

(d) Usage of modifier *near*

Table 6: Usage of modifiers *and*, *or*, *not* and *near*

"close to" or *"near"* itself. Generally, geographical concepts or relations are essential to provide a high-quality tourism information service. Comparing the modifier usage statistics a remarkable detail is noticeable. In 122 out of 1,425 queries (8,6%) the modifier *"near"* is used. This circumstance makes *"near"* to the modifier second-most frequently used, in the queries collected during the field trial. A common way to use *"near"* is to find accommodations in the surroundings of popular sites, cities or facilities, e.g. *"I am looking for a hotel with sauna and pool in St. Anton near the Galzig-Seilbahn"*.

Table 7 (a) illustrates the combined usage of the modifiers *"and"* and *"or"*. Most commonly used is a combination of one *"or"* and several *"and"* modifiers, e.g. two *"and"* and one *"or"* are used in 17 German queries. As shown in Table 7 (b), the usage of *"near"* corresponds with the presence of an *"and"* modifier.

We can say that the sentence complexity, i.e. the frequency of concept combination, is relatively low. In general, queries are formulated on the basis of combining concepts in a simple manner, e.g. *"I am looking for a room with sauna and steam bath in Kirchberg"*. Only a small subset of queries consist of complex sentence constructs that would require a more sophisticated sentence evaluation process. For instance, if the scope or type of the modifier cannot be determined correctly. As an example, consider the query *"I am looking for an accommodation in Serfaus, Fiss or Ladis"*. For the reader who is not familiar with the geography of Austria, in particular the Tyrol in this case, we shall note that *Serfaus*, *Fiss*, and *Ladis* are names of towns, and collectively they refer to an attractive skiing resort. In contrast to the assumption that the default operator of combining concepts is *"and"*, the modifier *"or"* must be used to combine the geographical concepts in this sample query.

The fact that the level of sentence complexity is not very high suggests, that shallow text parsing should be sufficient to analyze the queries emerging in a limited domain like tourism. Nevertheless, we found out that regions or local attractions are important informations that have to be integrated in such systems. We also noticed that users' queries contained vague or highly subjective criteria like *"romantic"*, *"cheap"* or *"within walking distance to"*. Even *"wellness"*, a label broadly used in tourism nowadays, is far from being exactly defined. These concepts are difficult to model in the knowledge base of information systems and pose a challenge for the future.

		query language		
and	**or**	**german**	**english**	**totals**
1	1	9	1	**10**
	2	3	0	**3**
2	1	17	2	**19**
	2	3	0	**3**
3	1	16	0	**16**
	2	5	0	**5**
	3	2	1	**3**
	6	1	0	**1**
4	1	12	1	**13**
	2	3	0	**3**
	12	3	0	**3**
	16	1	0	**1**
5	1	8	0	**8**
	2	1	1	**2**
	3	2	0	**2**
6	1	2	0	**2**
	2	2	0	**2**
	3	2	0	**2**
7	1	2	0	**2**
8	1	1	0	**1**
	2	1	0	**1**
totals		**96**	**6**	**102**

(a) Combined usage of modifiers *and* and *or*

		query language		
and	**near**	**german**	**english**	**totals**
1	1	18	1	**19**
2	1	32	2	**34**
3	1	26	1	**27**
4	1	21	4	**25**
5	1	7	0	**7**
	2	0	1	**1**
6	1	2	1	**3**
7	1	2	0	**2**
8	1	1	0	**1**
totals		**109**	**10**	**119**

(b) Combined usage of modifiers *and* and *near*

Table 7: Combined usage of modifiers

5 Conclusions

Web-based tourism information systems are faced with a highly inhomogeneous mix of potential customers. The reason, obviously, has to do with the tourism domain because, pragmatically speaking, almost everybody is a tourist sometimes. Hence, people with highly different backgrounds regarding their language, their preciseness in the description of information needs, or their computer literacy, to name but a few, are the customers of a web-based tourism information system. To cope with this situation we designed a multilingual natural language query interface for *Tiscover*, the largest Austrian tourism platform. By way of this interface, the user can search for more than 13,000 accommodations in about 2,000 towns throughout Austria.

In this paper we have discussed the findings after a 10 day field trial where we collected about 1,400 queries, most of which in German language. Most importantly, the users are willing to type natural language queries to express their information needs. This observation is approved by a comparison with web-search engines, where the average number of words per query is substantially smaller than with our tourism information system. Second, the complexity of these queries is higher than with standard web-search engines. We have shown the distribution of various modifier combinations extracted from the queries. Third, our expectation that shallow language processing is sufficient given a limited application domain is backed by the fact that most of the query concepts which had their counterpart in the knowledge base where successfully extracted from the natural language query. Fourth, by way of this field trial allowing natural language descriptions of information needs as opposed to the strictly limited variability of tabular-based information entry, we have got an impression of what the customers actually look for. Among the most important things we just mention geographic information as when you describe the location of your preferred accommodation relative to some geographical landmarks. This gives enough room for interesting future research to improve the knowledge base of the system and thus to better serve the customers.

From the data collected during the field trial, we are confident that natural language query interfaces represent an attractive alternative interaction paradigm for e-commerce systems. This is especially true in case of inhomogeneous user groups as with tourism information systems. Such interfaces can assist substantially in reducing the distance between the available information and the potential customer. Moreover, the fear on the side of the service providers that only elaborate and hence costly natural language analysis techniques are applicable can be reduced because of our observations during the field trial. In general, the complexity of the sentences expressing the user's information need is tractable with shallow language processing techniques as in our application.

Acknowledgments

Thanks are due to the people at *Tiscover*, especially Andreas Bacher, Claudia Bär, Reinhard Hauer, and Konrad Plankensteiner, for providing the data and for the good cooperation in carrying out the field test.

References

[1] I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. Natural language interfaces to databases – an introduction. *Journal of Natural Language Engineering*, 1(1):29–81, 1995.

[2] H. Berger, M. Dittenbach, D. Merkl, and W. Winiwarter. Providing multilingual natural language access to tourism data. In *Proc. of the 3rd International Conference on Information Integration and Web-based Applications and Services (IIWAS 2001)*, Linz, Austria, September 10-12 2001.

[3] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proc. of the 3rd Int'l Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, Las Vegas, NV, 1994.

[4] R. Dale, H. Moisl, and H. Somers, editors. *Handbook of Natural Language Processing*. Marcel Dekker, New York, NY, 2000.

[5] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.

[6] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ, 2000.

[7] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[8] J. Nielsen. Noncommand user interfaces. *Communications of the ACM*, 36(4):83–99, April 1993.

[9] J. Nielsen. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing, 2000.

[10] P. O'Brien. Dynamic travel itinerary management: The ubiquitous travel agent. In *Proc. of the 12th Australasian Conference on Information Systems*, Coffs Harbour, Australia, 2001.

[11] B. Pröll, W. Retschitzegger, R. R. Wagner, and A. Ebner. Beyond traditional tourism information systems – TIScover. *Information Technology and Tourism*, 1, 1998.

[12] B. Shneiderman, D. Byrd, and W. B. Croft. Sorting out searching. *Communications of the ACM*, 41(4):95 – 98, April 1998.

[13] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large AltaVista query log. Technical Report 1998-014, digital Systems Research Center, 1998.