Understanding and supporting information seeking tasks across multiple sources
Amin, A.K.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Supporting Query Formulation with Thesaurus-based Autocompletion Interfaces

We investigate interfaces that assist a specific activity of information seeking, namely query formulation (research question 3). We carry out two user studies on thesaurus-based autocompletion. In the first study, we explore different strategies for grouping autocompletion suggestions using two different thesauri. The aim of the first study is to see which Group strategy is preferred and better understood by lay users. The results suggest that the best strategies depend on the different thesaurus structures. Group organization seems to be more appropriate for geographical names thesaurus (a domain specific thesaurus with symmetrical hierarchy), while Alphabetical works better for object names from a lexical database of the English language (a thesaurus with less symmetrical hierarchy covering a broad scope). In the follow-up study, we compare three different organization strategies (Alphabetical, Group and Composite) for location name search tasks. The aim of this study is to investigate the different organization strategies support users in query formulation. Based on this study, we derive guidelines to design thesaurus-based autocompletion interfaces (research question 2). The results of the second study indicate that for TGN thesaurus, Group and Composite organizations help users search faster, and are perceived easier to understand and to use than Alphabetical.

This chapter was published as “Organizing Suggestions in Thesaurus-based Autocompletion Interfaces” in the 31st European Conference on Information Retrieval (ECIR’09) conference (Amin et al. 2009) and was co-authored by Michiel Hildebrand, Jacco van Ossenbruggen and Lynda Hardman.
4.1 Introduction

There is a lot of interest in the Information Retrieval community for interactive query expansion features that can help improve user search performance and the quality of queries submitted. There are two types of interactive query expansion: real-time query expansion (provide suggestions during query formulation) and post-query expansion (provide suggestions after query formulation). Between the two types, real-time query expansion (RTQE), such as autocompletion has been most adopted in many operational search applications e.g. Google Suggest or Yahoo! Search Assist. RTQE is an attractive feature because it can reduce the required number of keystrokes, decrease the user’s cognitive load through term recognition (rather than recall) and help the user avoid typing errors (Hendrickson 1989; Jakobsson 1986). It can improve the quality of initial query for known-item as well as exploratory tasks (White and Marchionini 2006; White and Marchionini 2007). RTQE is better than post-query expansion because it lowers task completion time, increases search engagement and increases uptake of interactive query expansion (White and Marchionini 2006). Most research efforts are directed towards improving query expansion suggestions, e.g. (Agichtein et al. 2006; Efthimiadis 2000; Radlinski 2005; White and Morris 2007), and tend to pay less attention to the interface issues. Many RTQE use only list organization as presentation style. Prior work has led us to believe that different types of implementation of RTQE presentation would likely result in different user search performance. In (Beaulieu 1997), three different interfaces to the same retrieval system were compared. The study suggests that the quality and effectiveness of search depend on the combination of the retrieval system and its interface in supporting query expansion. Joho et al. (Joho et al. 2002) studied different query expansion presentation styles. They compared two types of organization strategies: alphabetically ordered list and menu hierarchy IQE interfaces. They found that even though there is no significant difference in the precision-recall between using the two interfaces, people finished the search task significantly faster when using the menu hierarchy. Another study (Joho et al. 2004) compared two different hierarchical IQE systems (based on the subsumption approach and trigger phrased on parent-child description) against a baseline (no suggestion). They found that accessing the hierarchies reduces search iterations, reduces paging actions and increases the chance to find relevant items. In practice, many variants of RTQE organization strategies have been deployed, such as:

- **List** organization strategy, such as by alphabetical list (WikiSearch), by past user query in (Kelly et al. 2010), by popular query/destination in (White et al. 2007).
- **Group** organization strategy, such as Google Suggest uses 2 groups: personal
history query and popular prefix match, or Naver.com uses 2 groups: popular prefix and suffix match.

• **Menu** organization strategy, such as a cascaded menu hierarchy in (Joho et al. 2002; Joho et al. 2004).

In this research, we focus on the presentation aspects of an autocompletion, namely organization strategies and how they influence users search performance. We are motivated by the usage of relationships of terms from a thesaurus to improve RTQE presentation. Certain relationships between terms from a thesaurus has been known to improve the quality of query expansion. Efthimiadis et al. (Efthimiadis 2000) investigated the terms used in a IQE for the INSPEC database. They reported that variants (synonym) and alternative terms (i.e. narrower, boarder and related terms) relationships are useful for query expansion. Similarly, Joho et al. (Joho et al. 2002) in their research found that for WordNet, the most useful relationships are hyponym, hypernym and synonym. Additionally, they also mention conceptual relation (e.g. teeth-dentist) is a meaningful relationship for query expansion. In this study, we investigate how to improve selection of terms in autocompletion interfaces. In particular, we explore the potential of hierarchical relations in thesauri to improve the organization of autocompletion suggestions. By imposing grouping and ordering strategies we provide a means of navigating the suggestions and finding the right terms faster and easier. We restrict our discussion to interfaces that syntactically complete the input based on exact or partial literal match. We do not consider query recommendation which tries to extrapolate queries based on certain (semantic) relations or algorithms, such as in (Bast et al. 2007). Semantic relations are used, if at all, primarily to organize the suggestions in the interface. We carry out two related studies. The first examines the quality of grouping strategies for different thesauri, the second investigates to what extent grouping and (alphabetical) ordering are able to influence the suggestion selection process.

### 4.2 Organization of Suggestions

In this section, we discuss the look and feel of different organization strategies for autocompletion suggestions used in Study 1 and Study 2 (see Fig.4.1 and Fig. 4.2) Examples are taken for TGN $^1$ autocompletions, similar visualization and algorithm is applied to WordNet.

**Alphabetical order** — Fig.4.1a shows autocompletion suggestion in an alphabetical order. The suggestions are organized in the following priority: prefix match

---

$^1$Thesaurus for Geographical Names [http://www.getty.edu/research/conducting_research/vocabularies/tgn/]
on primary literal (location name), prefix match on secondary literal (country name), e.g. the suggestion “Paris, Canada” is shown before “Paris, France”. Exact matches are presented first, followed by partial matches. The first part of suggestion consists of 15 items. When the user selects the “see more” button, all suggestion is presented as a long list.

**Group** — An organization that combines similar suggestions under a common heading. The grouping category is conveyed visually as a group title. Where terms are related by explicit thesaurus relations, any of these relations can be used as a basis for grouping. Grouping can be based on variants of hyponym relations. There are 2 types of grouping: predefined and dynamic. In predefined grouping the category is always of the same type. For example, TGN’s hierarchy is based on geographical containment (e.g. *Europe* > *France* > *Paris*). Grouping can be based on any predefined level within this hierarchy, e.g. grouping by country (Fig. 4.1b). Alternatively, a predefined category can be based on a common property, such as place type (Fig. 4.1c) e.g. inhabited place (city, village) or body of water (stream, lake).

Another variant is the dynamic grouping where the group heading differs and is determined by an algorithm that optimizes groups based on the number of suggestions retrieved. The desired groups can be preset taken from the top level hierarchy (Dynamic TB) or taken from the lowest (leaf) level hierarchy (Dynamic BU). Dynamic groups could provide an alternative grouping for thesauri with irregular hierarchical structures such as WordNet. Fig. 4.1d is an example of Dynamic TB group implemented for TGN.

**Composite** — A composite organization resembles a two level cascaded menu
hierarchy. It is groups similar suggestions into a single item (primary menu),
deferring their display to a secondary menu. Fig. 4.2d shows an example composite
suggestion interface, the primary menu contains all exact match of all location
names from the same country. The secondary menu displays more information of
the location names that allows disambiguation e.g. Kingston (the city) or Kingston
(the parish). This strategy retains the simplicity of alphabetical order, but shows
larger numbers of alternatives in the limited amount of screen real estate available.

4.3 User Studies

We conducted two user studies to investigate the benefits and trade-offs of differ-
ent strategies for organizing suggestions in autocompletion interfaces. The first
study was an exploratory study to investigate the effects of grouping strategies on
two different thesauri. The second study built on the result that grouping terms
in a thesaurus of locations can be beneficial. The study investigated different
organization strategies on the same set of suggestions.

**Technology** — The study was done using our autocompletion widget\(^2\) that works
on all major browsers supporting (X)HTML, CSS and Javascript. The client side
widget is an extension of the Yahoo! User Interface autocompletion widget (YUI
v. 2.3.1\(^3\)). The suggestion server has been implemented using SWI-Prolog’s web
infrastructure\(^4\). The autocompletion architecture is fully described in (Hildebrand
et al. 2007).

4.3.1 Study 1: Grouping Strategies

In Study 1, we investigate different variations of grouping as a type of suggestion
organization. As mentioned in the previous section, there are many alternative
implementations of grouping strategy using various term relationships in a thes-
aurus. The goal of Study 1 is to investigate to what extent grouping strategies for
autocompletion suggestions can be applied to thesauri and if so, which grouping
strategies are meaningful for users. We have chosen to implement similar group-
ing strategies for two different thesauri: a domain-specific thesaurus, TGN and a
global thesaurus, WordNet. Our intention was not to compare the two thesauri,
but to evaluate the suitability of different group strategy when implemented for
these thesauri. Our research questions in Study 1 are: Can group organization
strategy be implemented for the thesauri? Which group strategy is perceived the
best by users?

\(^2\)Demo is available at http://slashfacet.semanticweb.org/autocomplete/demos/
\(^3\)http://developer.yahoo.com/yui/autocomplete/
\(^4\)http://www.swi-prolog.org/
Interfaces — We selected 4 autocompletion interfaces to compare with each other for TGN and similarly 4 for WordNet. The four chosen interfaces for each thesaurus are those which we thought were best to offer to users after informal trials of different algorithms and combinations. For TGN, the grouping strategies are: grouping by country (Fig. 4.1b), grouping by place type (Fig. 4.1c) and Dynamic TB grouping (Fig. 4.1d). As a baseline, we used Alphabetical ordering (Fig. 4.1a). We refer these interfaces as the Location Name (LN) interfaces.

For WordNet, the 3 grouping strategies are: predefined grouping using the top nine WordNet category nouns from the hypernym hierarchy, and two dynamic groupings: Dynamic TB and Dynamic BU. Similarly, the Alphabetical order was chosen as a baseline. We refer these interfaces as the Object name (ON) interfaces.

Participants — Participants were recruited by sending out invitations to universities and research institutes from diverse departments, such as computer science, engineering and natural science. In total, 47 people responded. Participants were mostly students and some university employees. All participants reported that they use the Internet daily and are familiar with autocompletion interfaces (e.g. in email clients, search engines and web browsers), 14 participants have experience with autocompletion interfaces in specialized applications such as script editors and interactive script interpreters.

Procedure and Tasks — The study was done as an online interactive experiment. All session activities are logged. Prior to the task, participants answered a short questionnaire focusing about their experiences with autocompletion. Afterwards, every participant was assign tasks with 4 four TGN-LN interfaces (within subject design). For every LN, participants were given the same tasks: to formulate several location queries, such as Berlin (city name) or Alps (mountain system)), and find the correct location names from the suggestions presented in the interface. Afterwards they were encouraged to try out their own example queries and explore the interface responses. After completing the tasks, participants were asked to answered assessment questions about the quality of the groupings and to give their comments. Finally, participants were asked to rank their preferred strategy for LN, from the most to the least preferred, and provide reasons for their decisions. The same task and procedure were repeated by the participants for the WordNet-ON interfaces. Participants were asked to formulate object queries, such as Barbecue or Party, and answer the assessment questions about the quality of grouping in this interface. The assessment questions on the quality of the group organization were derived from criteria taken from the literature (Gonzales 1994; Hendrickson 1989; Hodgson and Ruth 1985; Lee and Raymond 1993; Rosch 1988). The answers to the assessment questions are given in a 7-Likert scale (1:low, 7:high). These criteria are:

Q1 - perceived similarity of items within the same group; “I think the items be-
longing to each group in this type of list are similar to each other."
Q2 - perceived difference of items between groups; “I think the items belonging to different groups in this type of list are different from each other.”
Q3 - affinity item and group title; “I think the relationship between the items and group title is clear in this type of list.”
Q4 - reasonable number of groups; “I think the number of groups in this type of list is appropriate.”
Q5 - group title appropriateness; “I think the titles of the groups in this type of list are clear.”

The order of the interfaces were counter balanced using the Latin Square scheme among the participants. Pilot sessions were conducted to ensure that the participants could perform the tasks and understood the questions. The time to complete the study was approximately 30 minutes.

**Results** — The data we collected from the experiment were processed qualitatively and quantitatively. Our server log indicates that in addition to trying all provided examples, additionally some participants explored the behavior of autocompletion interfaces by trying out their own examples, such as different cities, countries or river names (e.g. Rhein) for LN and various object names (e.g. muscle, mobile, partner) for ON. It is important for us to confirm that participants explore the behavior of the autocompletion beyond the given task before assessing the quality of the interfaces.

<table>
<thead>
<tr>
<th>Table 4.1: Top: Assessment scores (n=47 people, Study 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TGN-LN</strong></td>
</tr>
<tr>
<td>Question</td>
</tr>
<tr>
<td>Q1</td>
</tr>
<tr>
<td>Q2</td>
</tr>
<tr>
<td>Q3</td>
</tr>
<tr>
<td>Q4</td>
</tr>
<tr>
<td>Q5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>WordNet-ON</strong></th>
<th>**Mean Score (SD) *</th>
<th><strong>p-value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Predefined</td>
<td>Dynamic TD</td>
</tr>
<tr>
<td>Q1</td>
<td>4.19(1.56)</td>
<td>4.21(1.85)</td>
</tr>
<tr>
<td>Q2</td>
<td>4.64(1.47)</td>
<td>4.43(1.60)</td>
</tr>
<tr>
<td>Q3</td>
<td>4.13(1.81)</td>
<td>4.28(1.75)</td>
</tr>
<tr>
<td>Q4</td>
<td>4.19(1.72)</td>
<td>3.47(1.73)</td>
</tr>
<tr>
<td>Q5</td>
<td>3.83(1.81)</td>
<td>4.04(1.71)</td>
</tr>
</tbody>
</table>

* 7-Likert scale, score 1: strongly disagree, 7: strongly agree
Table 4.2: Bottom: Preferred grouping strategy (n=47 people, Study 1).

<table>
<thead>
<tr>
<th>TGN (LN)</th>
<th>Ease of use (SD)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place type</td>
<td>2.23(1.15)</td>
<td>.16</td>
</tr>
<tr>
<td>Dynamic TD</td>
<td>2.35(1.09)</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>2.67(1.13)</td>
<td></td>
</tr>
<tr>
<td>Alphabetic</td>
<td>2.74(1.09)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WordNet (ON)</th>
<th>Ease of use (SD)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabetic</td>
<td>1.98(1.23)</td>
<td>.02</td>
</tr>
<tr>
<td>Dynamic TD</td>
<td>2.62(.97)</td>
<td></td>
</tr>
<tr>
<td>Predefined</td>
<td>2.68(1.09)</td>
<td></td>
</tr>
<tr>
<td>Dynamic BU</td>
<td>2.72(1.06)</td>
<td></td>
</tr>
</tbody>
</table>

- **Assessment:** The participants’ assessments for six grouping strategies are shown in Table 4.1. We examine each question to understand the characteristics of each grouping strategy using Friedman two-way analysis by ranks\(^5\). For LN we found that: (a) Place type grouping scored best with respect to perceived similarity - Q1 ($\chi^2(2)=7.36$, $p=\ldots.03$)\(^6\) (b) Country grouping scored best with respect to group title appropriateness - Q5 ($\chi^2(2)=6.77$, $p=\ldots.03$)\(^7\) (c) Country grouping scored lowest with respect to the number of groups - Q4 ($\chi^2(2)=8.11$, $p=\ldots.02$)\(^8\) Perceived similarity indicates the cohesiveness between the suggestions in a group. Place type grouping scores highest for this aspect. Alternatively, the Country group strategy gives most representative group titles (Q5) but scores poor on the number of groups (Q4). One disadvantage of our implementation for the Country group strategy is that we did not make any limitation on the number of groups allowed. Because of this, the autocompletion list can potentially be very long. This is an adjustable parameter of the interface and does not depend on the characteristic of the thesaurus. The assessment score indicates that from the 3 types of LN grouping, Country and Place type are relatively good grouping strategies that each excel in different qualities.

For the ON interfaces, we found that: (a) Dynamic BU group scored lowest with respect to perceived difference - Q2 ($\chi^2(2)=10.17, p=\ldots.01$)\(^9\) (b) Dynamic TB group

---

\(^{5}\)Nonparametric statistics is used as the data did not meet parametric assumptions

\(^{6}\)Wilcoxon signed ranks (WSR) post-hoc test result for Q1: Place type scored sig. higher than Dynamic TB ($p<.05$).

\(^{7}\)WSR post-hoc test result for Q5: Country scored sig. higher than Dynamic TB ($p<.05$) and Place type ($p=.03$)

\(^{8}\)WSR post-hoc test result for Q4: Country scored sig. lower than Dynamic TB ($p=.02$) and Place type ($p=.01$)

\(^{9}\)WSR post-hoc test result for Q2: Dynamic BU scored sig. lower than Predefined ($p=.01$)
scored lowest with respect to the number of groups - Q4 ($\chi^2(2)=9.66$, $p=.01$)\textsuperscript{10} The results showed that none of the ON group strategies excels from each other in the assessment score. We only found that the Dynamic TB and Dynamic BU groups perform the worst in Q2 and Q4. We think this is because the dynamic group strategies actually add to participant’s cognitive load when they are trying to go through the suggestion list. No grouping strategy in ON is assessed the best by our participants. The reason for this will be clear in the next results where we compared all group strategies against a baseline (Alphabetical order) and examine users preferences.

- Preference: Table 4.2 shows the Mean Rank of each grouping strategy for LN and ON. A low Mean Rank score indicates most preferred, and a high score is least preferred. Using the Friedman two-way analysis by ranks, we found that there is no strong preference in any of the location grouping strategies ($\chi^2(3)=5.14$, $p>.05$). From the comments provided by the participants, we see that participants prefer different interfaces for different reasons. We conducted the same analysis for the four ON interfaces and found a different result. Participants strongly preferred the Alphabetical order to all other organization strategies ($\chi^2(3)=10.38$, $p=.02$)\textsuperscript{11} From the participants’ comments, we understood that they found it difficult to understand the ON grouping strategies. This could explain the strong preference for Alphabetical order.

- Comments: Participants’ comments gave us an explanation as to their assessment decisions and preferences. It seems that most decisions on choosing a LN interface is based on personal preference. “By country seems more logical and pragmatic. Place type takes some getting used to but could work fine. Dynamic (grouping) gets confusing, Alphabetical (list is) not very clarifying.” [P6]. For the ON interfaces, participants opinion are more uniform. The main comment was that many participants struggle with understanding ON grouping strategies. The baseline (Alphabetical order) seems to be the easiest to understand based on their past experience with finding terms in a dictionary. “...I am more familiar with encyclopedic or dictionary structuring. The problem with such group autocompletion advice is that the adaptation process is quite time costly.” [P25].

Retrospective — The main goal of study 1 was to get a feel for how users perceive different grouping strategies. More precisely, we want to find out if and how the different structures of the thesauri used effect the user’s perception, and whether the resulting groupings make sense at all.

Ideally, the best grouping strategies are the ones that scores highest on all five

\textsuperscript{10}WSR post-hoc test result for Q4: Dynamic TB scored sig. lower than Predefined ($p<.05$) and Dynamic BU ($p=.03$)

\textsuperscript{11}WSR post-hoc test result for Mean Rank of preference: Alphabetical scored sig. lowest (i.e. strongly preferred) then Predefined ($p=.02$), Dynamic TB ($p=.04$) and Dynamic BU ($p=.01$).
assessment scores (Table 4.1) and most preferred (Table 4.2). However, this is not the case. For LN, we found that one grouping strategy is better in some aspect while others in another aspect. For ON, we did not found any favoured grouping strategy.

Thus, we concluded that grouping strategies may not be suitable for every type of thesaurus. For a domain-specific thesaurus, such as TGN we could find a sensible grouping strategy that people could understand relatively easily. In a global thesaurus such as WordNet, however, we found different results. The users preference, assessment scores and participants’ comments lead to the conclusion that for WordNet the group organization may not be the best strategy to use. In cases where the underlying thesaurus does not provide the information necessary for appropriate grouping, Alphabetical ordering is the best option.

4.3.2 Study 2: Organization Strategies

Based on what we have learned in Study 1, we conducted a follow up study. We narrowed down the scope of study 2 by only investigating autocompletion for TGN. We decided not use WordNet since none of our group strategies for WordNet outperformed the baseline (Alphabetical). The goal of the second study is to compare three types of autocompletion: Alphabetical order, Group and Composite. We would like to investigate which interface helps users to search for terms from a thesaurus the fastest and easiest. To be able to come to this conclusion, we setup an experiment where users are required to use autocompletion for known-item search tasks. To evaluate speed, we measure performance in time to complete task (objective measurement). To measure ease-of-use, we took three subjective measurements: user assessments, preferences and comments. Additionally, we analyze the quality of keywords provided in each condition.

**Interfaces** — In this study, we compared 4 different interfaces, namely: Alphabetical order (Fig.4.2b), Group (Fig.4.2c), Composite (Fig.4.2d) and a no autocompletion (NAC) interface (Fig.4.2a).

**Participants** — We recruited participants in the same manner as in the first study. In total, 41 people participated. Participants were aged between 16-66 years ($M=30.90$, $SD=10.45$). In general, participants use the Internet frequently ($M=34.96$, $SD=19.51$) (hours per week), and have medium to high familiarity with autocompletion interfaces.$^{12}$

**Procedure** — Each participant is assigned interfaces: NAC, Alphabetic, Group and Composite (within subject design). The order of the conditions were counter

---

$^{12}$: low familiarity, 5:high familiarity; Autocompletion in search engines ($M=3.40,SD=1.34$), email client ($M=4.12,SD=1.17$), address browser($M=3.86,SD=1.46$), Misc.: autocompletion in MS Visual Studio, Eclipse IDE
balanced using the Latin Square scheme among the participants. Pilot sessions were conducted to ensure that the participants could perform the tasks and understood the questions. The time to complete the study is approximately 25-30 minutes. In the experiment, participants started by answering general questions about their experience in using the Internet and autocompletion. Participants were then given a trial session to get accustomed to the interfaces. During the experiment, participants were given 24 tasks. In every task, time measurements were taken and participants were asked to assess the usability of the interface afterwards. We are interested in comparing the usability aspects of the different interfaces. After every interface, participants answered two questions (5-Likert scale):

Q1 - “I find this interface easy to use.”
Q2 - “I find the organization of the suggestions easy to understand.”

At the end, participants were asked to rank the autocompletion interfaces based on their preference and to give reasons for their choices.

Task — Participants were given 24 tasks (3 tasks per interface). To simulate a realistic search task, participants were asked to search and specify the birth place of a famous person (see Fig. 4.3). They were allowed to find the answers in Wikipedia and then fill in their answer using the autocompletion interface. Participants were encouraged to use autocompletion but could choose not to use it if they could not find the right suggestion from the list. We have chosen the non trivial tasks such as locations with exactly the same name. Thus, for all questions, the need for disambiguation and choosing the correct terms was clear. For example, the birth place of Kurt Kobain (Aberdeen, Washington, USA) has at least 56 other similar place name matches, of which Aberdeen in the UK will most likely be the most familiar to our European participants. The times recorded are the autocompletion typing time only. We disregard the time it takes for the participant to browse the Web and look for answers.

Results

- **Mean keystrokes:** As expected, an autocompletion interface reduces the number of keystrokes required to type. On average, users typed almost twice as many characters in the NAC condition compared to when using autocompletion (see Table 4.4). Additionally, we found that some participants copied and pasted the location name they found from Wikipedia. This behavior was identified from the interaction event log and is estimated to be about 7.5% from the total tasks performed by all participants.

- **Performance in Time:** Table 4.4 shows the mean time it took for participants to complete a task. This time constituted the time from the first keystroke typed until
selecting a suggestion (for the autocompletion conditions) or hitting the return key (for the NAC condition). When we compare the performance of the individual autocompletion interfaces, we find that Group and Composite are significantly faster (47% and 45% resp.) than the Alphabetical order. We conclude that both Group and Composite strategies help the user search for terms faster than Alphabetical order.

- **Quality of keywords**: Table 4.3 shows the quality of keywords provided by participants. The quality of keywords is measured by how accurately the location names are given. We found four levels of accuracy (from low to high): strings that consist of one term (mostly city names which are ambiguous because there exist many cities the same name, e.g. “Kingston”), strings that consist of two terms (mostly a combination of city and state, or city and country, e.g. “Kingston, USA”), strings that consist of three terms (mostly a combination of city, state and country names, e.g. “Kingston, Texas, USA”) and keywords which were correctly chosen from the

\[\text{\textsuperscript{13}}\text{WSR post-hoc test result for Time: Group is sig. faster than Alphabetical (}p<.05\text{). Additionally, Composite is sig. faster than Alphabetical (}p<.05\text{).}\]
4.3 Thesaurus-based Autocompletion Interfaces

Table 4.3: Quality of keywords provided by participants (492 tasks, 41 people, Study 2)

<table>
<thead>
<tr>
<th>Interface</th>
<th>NAC</th>
<th>Alphabetical</th>
<th>Group</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total correct keyword</td>
<td>96.7%</td>
<td>86.2%</td>
<td>95.1%</td>
<td>84.5%</td>
</tr>
<tr>
<td>a. Unique concept</td>
<td>n/a</td>
<td>77.2%</td>
<td>86.2%</td>
<td>82.9%</td>
</tr>
<tr>
<td>b. One term</td>
<td>14.6%</td>
<td>2.4%</td>
<td>0.8%</td>
<td>0%</td>
</tr>
<tr>
<td>c. Two terms</td>
<td>53.7%</td>
<td>6.5%</td>
<td>5.7%</td>
<td>0%</td>
</tr>
<tr>
<td>d. Three terms</td>
<td>28.4%</td>
<td>0%</td>
<td>2.4%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Total incorrect keyword</td>
<td>3.2%</td>
<td>13.8%</td>
<td>4.9%</td>
<td>15.4%</td>
</tr>
<tr>
<td>a. Select wrong item</td>
<td>n/a</td>
<td>13.0%</td>
<td>4.9%</td>
<td>15.4%</td>
</tr>
<tr>
<td>b. Typing error</td>
<td>2.4%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>c. No answer</td>
<td>0.8%</td>
<td>0.8%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 4.4: User search performance and preference (492 tasks, n=41 people, Study 2)

<table>
<thead>
<tr>
<th>Interface</th>
<th>NAC (SD)</th>
<th>Alphabetical (SD)</th>
<th>Group (SD)</th>
<th>Composite (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean no of keystrokes</td>
<td>19.20(6.86)</td>
<td>8.55(4.50)</td>
<td>7.89(4.81)</td>
<td>7.91(3.82)</td>
</tr>
<tr>
<td>Mean time in s</td>
<td>5.94(3.41)</td>
<td>38.93 (46.87)</td>
<td>18.36 (10.99)</td>
<td>17.62 (12.25)</td>
</tr>
<tr>
<td>Mean pref. rank</td>
<td>2.93(1.23)</td>
<td>2.71 (.90)</td>
<td>1.98 (1.11)</td>
<td>2.39 (1.02)</td>
</tr>
<tr>
<td>Mean score Q1 *</td>
<td>3.07(1.21)</td>
<td>2.59 (.87)</td>
<td>3.34 (1.39)</td>
<td>3.56 (.90)</td>
</tr>
<tr>
<td>Mean score Q2 *</td>
<td>n/a</td>
<td>3.05 (1.24)</td>
<td>3.73 (1.10)</td>
<td>3.61 (.95)</td>
</tr>
</tbody>
</table>

* 5-Likert scale, score 1:strongly disagree, 5:strongly agree

suggestions (unique concepts from the thesaurus). All autocompletion interfaces have a high percentage of correct keywords (all above 84.5%). The quality of keywords provided by participants, however, differs when using autocompletion and without. In NAC, most keywords consist of merely 2 terms (53.7%), which is in many cases not sufficient for location name disambiguation. For example, there are 47 places named Kingston in the USA. Only about a third of the cases in NAC (28.4%) consist of three terms. In contrast, keywords provided in the other autocompletion interfaces are mostly high quality keywords that are unique concepts (86.2% for Alphabetical, 95.1% for Group, and 84.5% for Composite). We also identified three types of incorrect keywords: incorrect terms selected from the autocompletion suggestions, typing errors and blank entries where participants gave no keyword at all. A closer look at the incorrect keywords provided by participants reveals that most errors in the NAC conditions are typing mistakes (2.4%), while most errors in the autocompletion interfaces are wrong autocompletion selection. For example, selecting Ottawa (the river) instead of Ottawa (the city). From all 3 autocompletion interfaces, Group organization generates least error (4.9%) compared to Composite (15.4%) and Alphabetical (13.8%). The results show that
even though NAC is slightly faster, the quality of keywords provided in any of the autocompletion interfaces are far higher.

- **Perceived ease-of-use:** We gathered participants assessments on the ease-of-use of each interface (see Table 4.4). In general, people find the Group and Composite interface easier to use than Alphabetical and NAC interface \( (\chi^2(3)=17.52, p\ll.05) \)\(^{14}\) In a follow up question (Q2), we wanted to know specifically if people understood the organization strategy. Most people agree with the statement that Group and Composite suggestion organization is easier to understand than Alphabetical list \( (\chi^2(2)=8.12, p=.02) \)\(^{15}\). We conclude that both Group and Composite interfaces are perceived easier to use and understand than the Alphabetical order.

- **Preference:** Our analysis shows there is a preference for Group strategy (see Table 4.4), although Composite is not far behind \( (\chi^2(3)=12.6, p\ll.05) \)\(^{16}\). From the comments made by the participants we understand more about the reasons behind the users preference. Participants acknowledge that autocompletion suggestions help avoid typing mistakes and enable them to express more keywords than they would otherwise have thought of. “The lack of autocompletion choices prevents me to give a proper answer for question X.”[P1]. In general, participants think Group organization is better. “It’s comfortable to see the countries separated” [P2]. “You know where you have to go. You get a better overview” [P16]. For many participants, the Composite organization is relatively new. The main disadvantages of Composite are: a) requires more interaction with the interface before making a selection (e.g. mouse movement and click) and b) submenu interaction requires getting used to “took several seconds to discover the small arrows. After that, the interface is easy to use” [P4]

### 4.4 Discussion

**Study limitation** — In Study 1, the grouping strategies tested were developed partly by trial and error in combination with educated guesses. We came up with a range of possible grouping strategies, which we tested informally. We only formally tested the three different grouping strategies which performed best in the informal test. The result of Study 1 shows that, in contrary to the TGN grouping strategies, our best grouping strategies for WordNet were not helpful for users and people pre-

---

\(^{14}\) WSR post-hoc test result for Q1: Group is sig. perceived easier-to-use than Alphabetical \( (p\ll.05) \). Composite is sig. perceived easier-to-use then Alphabetical \( (p\ll.05) \). No difference between Group and Composite.

\(^{15}\) WSR post-hoc test result for Q2: Group organization is sig. perceived easier to understand than Alphabetical \( (p=.01) \). Composite organization is sig. perceived easier to understand than Alphabetical \( (p=.04) \). No difference between Group and Composite.

\(^{16}\) WSR post-hoc test result for preferred interface: Group organization is sig. preferred than Alphabetical \( (p\ll.05) \) and NAC \( (p\ll.05) \). No sig. difference between Group and Composite.
ferred Alphabetical ordering as an organization strategy. We acknowledge that it might be the case that we did not succeed to find the appropriate grouping for WordNet. Therefore it is reasonable to only conclude that for WordNet grouping strategies which were tested, none outperformed the Alphabetical list.

During the experiment, we observed that there are some cut and paste behavior specially for the NAC interface. This might provide additional explanation as to why participants complete NAC tasks faster than other tasks. We expect that if all participants were only allowed to type (not cut and paste), the autocompletion interfaces would show comparable time performance.

**Alphabetical order** — When using a global thesaurus, such as WordNet, Alphabetical order organization seems to be the best option. This organization requires very little learning effort. The downside of this organization is, as one participant points out, that it provides no “overview” when there are many suggestions.

**Grouping strategy** — We learned from the first study that a grouping strategy should be chosen carefully because not every grouping strategy is suitable to use. The TGN groupings produced by the thesaurus hierarchy seem to be more natural than WordNet groupings. In many of our pairwise statistical comparisons between Group and Composite organization, we found no significant differences. Based on this study alone, we cannot see a clear advantage one type of organization over the other. We can say, however, that the Group organization has a tendency to expand the length of suggestion interface vertically, whereas the Composite organization has a tendency to reduce the length of the suggestion interface. Therefore, depending on the thesaurus used and the length of suggestions it produces, the Composite organization might have an advantage.

**Improve autocompletion** — The server log indicates that some people use commas and make keystroke errors. We learned that in order to make a good autocompletion interface, there are a number of supporting functionalities that are indispensable: (a) Compensate for non alphanumeric letters such as white space(s) and commas. For example, the system should know that *Kingston Jamaica* is the same query as *Kingston, Jamaica*. Our finding is consistent with the study in (Teevan et al. 2007) on how people express similar queries in different ways. (b) Users may make typing mistakes (e.g. *Ottawa*, *Ottowa*, *Otawa*). Spell check and giving suggestions based on likely spelling would be a useful feature.

**Incentive to use autocompletion** — We observe that autocompletion can stimulate people to be more precise in their keywords. Even though participants were instructed to be "as precise as possible”, the keywords provided in the no-autocompletion tasks are largely ambiguous: only less than a third of the keywords consist of detailed information (i.e. city, state and country). This is in large contrast to the keywords provided by the participants when using autocompletion suggestions where they are mostly unambiguous location concepts. Autocomplet-
tion allows people to provide high quality keywords from which an information retrieval system can benefit. We believe that users are willing to spend more time to formulate a more elaborate query with the help of autocompletion interfaces if the option to use is made available.

4.5 Conclusions and Future work

We conducted serial user studies to compare different kinds of organization for autocompletion suggestions that can help improve known-item search task. In the first study, we found that grouping strategies might not be suitable when using a global thesauri, such as WordNet and only certain grouping strategies could be used for TGN. Based on what we have learned, we conducted a second study where we compared three different autocompletion suggestion interfaces. In general, we found that the quality of keywords provided by users are better with the autocompletion; Group and Composite organization help users search faster than when using the Alphabetical order; users perceive Group and Composite easier to use and to understand. We are currently integrating autocompletion with our applications and evaluate its performance for a domain-expert annotation task. In addition to this, we will improve autocompletion interface to detect similar query strings identified in (Teevan et al. 2007), such as synonyms, extra whitespace and word swaps.