Starting Conversations with Search Engines - Interfaces that Elicit Natural Language Queries

Andrea Papenmeier, Dagmar Kern, Daniel Hienert
firstname.lastname@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Alfred Sliwa, Ahmet Aker, Norbert Fuhr
firstname.lastname@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

ABSTRACT

Search systems on the Web rely on user input to generate relevant results. Since early information retrieval systems, users are trained to issue keyword searches and adapt to the language of the system. Recent research has shown that users often withhold detailed information about their initial information need, although they are able to express it in natural language. We therefore conduct a user study (N = 139) to investigate how four different design variants of search interfaces can encourage the user to reveal more information. Our results show that a chatbot-inspired search interface can increase the number of mentioned product attributes by 84% and promote natural language formulations by 139% in comparison to a standard search bar interface.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; **User studies**; *Graphical user interfaces*; **Empirical studies in visualization**; • **Information systems** → **Query intent**.

KEYWORDS

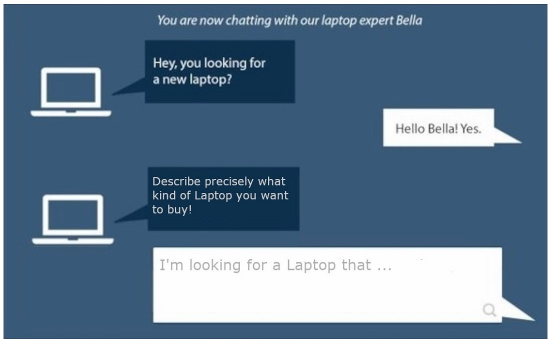
Information Need; Query Formulation; E-Commerce; User-Centered Design.

1 INTRODUCTION

When searching the Web with search engines, users have to communicate what they are looking for in a machine-interpretable manner. Search engines focusing on products often offer additional possibilities to navigate, such as filters or facets. However, research on product search engines shows that only a fraction of the actual information need is present in the initial search query [14]. A reason for this could be the vocabulary problem [9]: The structured data in the databases use a different vocabulary than the natural language of users. Hence, the user has to adapt to the system's language to maximise search success.

With voice assistants and voice search being on the rise, search engines need to process longer and more complex inputs. When equipped with appropriate capabilities to process such input, search engine performance can even be improved [8]. This opens up new opportunities: If users express their information need through natural language and disclose more information, search performance can be improved. However, keyword search is deeply ingrained in the users' minds [11].

In our research, we investigate how an interface can trigger users to give more information on their needs directly at the first interaction in a product search scenario. We design a user study to evaluate and compare four interfaces concerning the differences in query length, information content and formulation characteristics. Our key findings show that a chatbot-inspired interface (see Figure 1) succeeds in eliciting more information about the user's information need than traditional search bar interfaces, while also triggering a more natural query language.



You are now chatting with our laptop expert Bella

Hey, you looking for a new laptop?

Hello Bella! Yes.

Describe precisely what kind of Laptop you want to buy!

I'm looking for a Laptop that ...

Figure 1: Final design of the chatbot interface.

2 RELATED WORK

Although research showed that longer queries could produce better results for information seeking tasks, e.g. [6, 7], people usually tend to use short search queries [3]. There are many approaches to support users in finding relevant information, e.g. through facets, recommendations, implicit and explicit user feedback. However, only few works have tried to motivate users to type in more query terms and thus provide more detailed information about their initial information need. Belkin at al., for example, showed that a query-entry box with several lines led to longer queries than a line mode search bar [6] and that query lengths were significantly longer when the query box was labelled with "Information problem description (the more you say, the better the results are likely to be)" than when it was labelled with "Query terms" [7]. Furthermore, they found that longer queries significantly increase searchers' satisfaction [7]. In contrast, Agapie et al. [1] found that telling users that longer queries deliver better search results does not influence query length. However, they showed that using a coloured halo around the search bar motivates searchers to provide significantly more query terms in a complex Web search scenario. Hiemstra et al. [10] evaluated the proposed halo effect in a website search system in a 50-day A/B test (N = 3506) but could not confirm the positive impact on query length. They conclude that this approach might be sensitive to the search task and search context. Kelly and Fu [12] show that additional information (domain knowledge, the information need, and search motivation) help to increase the retrieval performance. Likewise, Bendersky, Croft and Bruce [8] propose a machine learning method to extract the key facts from long queries. Their system performs better on longer natural language queries as compared to shorter, keyword-like queries.

A reliable information need elicitation is getting more critical with the increasing use of voice assistant systems. Without a graphical user interface, refining the search via facets and exploring the results and recommendation lists becomes cumbersome. Research in the context of conversational search has explored asking clarifying questions [2] or coached conversational preference elicitation [15]. With one good question, Aliannejadi et al. [2] improved the retrieval performance by over 150%. Focusing on conversations, however, requires processing natural language (with challenges such as vague language and ambiguity [4]), which is, so far, not supported by common product search engines.

3 USER STUDY

Previous research has explored changes on the graphical user interface and in the interaction design to elicit longer search queries. User studies, so far, have not focused on the generation of natural language queries to support elicitation of the natural and complete information need, especially not in the context of product search. We close this gap by investigating the following research questions:

RQ 1: Do users reveal more about their information need when interacting with more conversation-like interfaces?

RQ 2: Are users more inclined to use natural language when interacting with more conversation-like interfaces?

To answer the research questions, we design interfaces based on cues from prior literature and evaluate them in an online study with a between-subjects design. The interface designs, questionnaire, and annotation guidelines are available online[1].

3.1 Iterative Interface Design

Initially, seven interfaces were designed that implemented either different search bar sizes (inspired by [6]), a dialogue-inspired speech bubble and chatbot design inspired by conversational search, or different avatars (inspired by agents in e-commerce [13]). All seven interfaces were tested in a between-subject online pilot study with 60 participants. Besides issuing a query, the participants gave qualitative feedback about their experiences. Finally, the set of interfaces was reduced to four interfaces that showed the most diversity in query formulations. Interfaces using avatars were excluded to focus on search bars and dialogue-inspired designs: an interface with a small search bar (I1), one with a big search bar (I2), one with a direct question in a speech bubble (I3), and a chatbot (I4, see Figure 1). As placeholders (or the absence of placeholders) impact the query formulation, we use the exact same placeholder text for all interfaces ("I'm looking for a laptop that..."). Hence, effects of the placeholder impact all conditions equally.

3.2 Scenario and Procedure

To evaluate the interfaces, we set up an online study on SoSci Survey[2]. After giving consent, participants receive the scenario and task description (adapted from [5, 14]):

> *Your laptop broke down yesterday. Today, you are searching for a new device. You decide to search online. You find the following website. Please use the search bar in the screenshot to search for your desired laptop.*

For I1, I2, and I3, participants submit their query and are redirected to a dummy result page to complete the search experience. For I4, participants submit the initial query, but then receive a generic follow-up question from the chatbot, which asks them to give more details. This design is chosen to simulate a chatot-like interaction. After the second prompt, participants of I4 also continue on the dummy result page. Subsequently, three open questions about their experience are asked, as well as closed questions about the individual domain knowledge and demographic background.

3.3 Measures and Analysis

To investigate whether the queries differ in the informational content (**RQ 1**), we analyse the submitted search query with the following measures: (a) count of words, (b) count of key facts – key facts being descriptions of product attributes, e.g. "i5 processor", "large screen" –, and (c) the attribute group per key fact (e.g. "processor", "screen"). For analysing differences in formulations concerning natural language (**RQ 2**), we determine per query: (d) count of vague words – ambiguous words that cannot clearly be mapped onto product attributes or values e.g. "good", "decent" –, (e) occurrences of grammatical word types (part of speech), and (f) sentence completeness (scale of 0-2, with 0 = keywords or bullet points, 1 = partial sentences, and 2 = full sentences). We use the NLTK[3] Python package for automatically retrieving (a) and (e). For extracting (b), (c), (d) and (f), we manually annotate the queries with two annotators who discuss discrepancies until a final annotation is found.

Finally, to test for significant differences between a group of independent samples (e.g. comparing conditions), we use the Kruskal-Wallis test with a pairwise Mann-Whitney U-test as post-hoc analysis, applying the Bonferroni correction to account for the multiple testing bias. For analysing correlations, we use the measure of Spearman's rho.

3.4 Participants

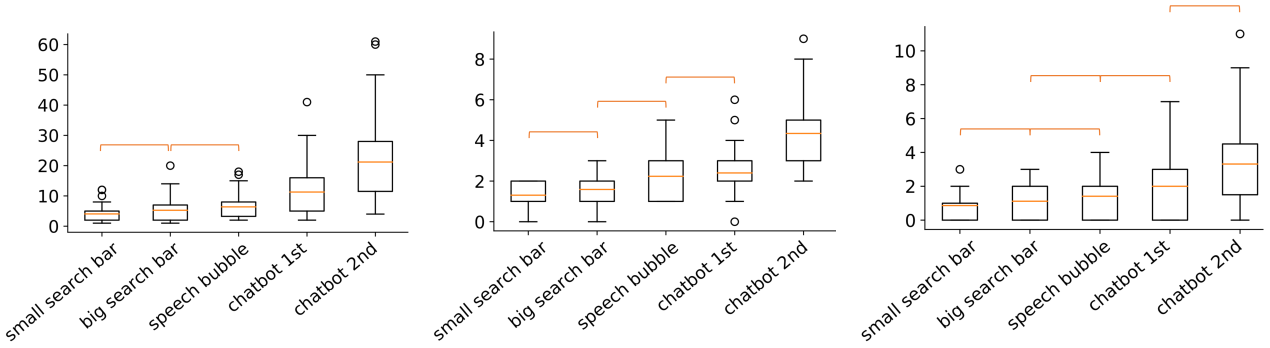
Overall, we recruited 139 participants (57 male, 80 female, 2 diverse) on the online crowdsourcing platform Prolific[4]. Participants were evenly spread over the four interfaces (N = 36, 34, 34, 35). As the sample group per condition is rather small, we aimed for a homogeneous sample: Users had to be residents of the US, the UK, or Ireland, native English speakers, and had to be "digital native" adults (ages 18-40, M = 28.3, STD = 5.9).

[1] https://git.gesis.org/papenmaa/chiir21_naturallanguageinterfaces

[2] https://www.soscisurvey.de/en/index

[3] https://www.nltk.org

[4] www.prolific.co



(a) query length (words)

(b) keyfact count

(c) vague word count

Figure 2: Boxplots of query length (a), key fact count (b), and vague words (c). Orange brackets signify an absence of significant difference between conditions, no bracket means significant difference between conditions.

4 RESULTS

4.1 Information Content

To investigate whether users reveal a different amount of information about their preferred product across conditions, we analyse the queries concerning their length, number of individual key facts, and the attribute groups. An exemplary query issued in I3 is given in the following, with key facts marked in bold:

"I'd like something ***lightweight***, ***easy to use*** *with a* ***long battery life***

The initial queries of I4 are significantly longer (M = 11, STD = 9) and contain significantly more key facts (M = 2.4, STD = 1.2) than both search bar interfaces I1 and I2 (see Figures 2a and 2b). On average, I4 leads to 84% more key facts than I1. A generic follow-up question can further achieve a significant raise in length (M = 21, STD = 15) and key facts (M = 4.3, STD = 1.7). We found a moderately positive correlation between domain knowledge and query length (r = .41). A difference in domain knowledge could bias the results if the average domain knowledge differed across conditions. A Kruskal-Wallis test, however, did not show any difference in domain knowledge (p = .564). Although participants with more domain knowledge tend to write longer queries, there is no correlation between domain knowledge and the number of key facts in a query (r = .01). However, comparing the length and the number of key facts, we find a strong positive correlation (r = .67).

We clustered all key facts into groups according to the product attribute they were addressing. Overall, we identified 26 unique product attributes, of which 13 appeared in queries of all conditions: brand, model, memory, graphics card, RAM, screen, battery, size, software, price, performance, quality, and purpose. In queries of I1, we found 15 unique attributes, 17 in I2, 22 in I3, 20 in I4, and 23 in I4 after the follow-up question. Attributes that were only mentioned in a dialogue-inspired condition (but not in a search bar condition) were, for example, the design of a laptop, the keyboard or the usability. This shows that the dialogue-inspired designs (I3, I4) elicited not only more key facts but also a greater variety of key facts.

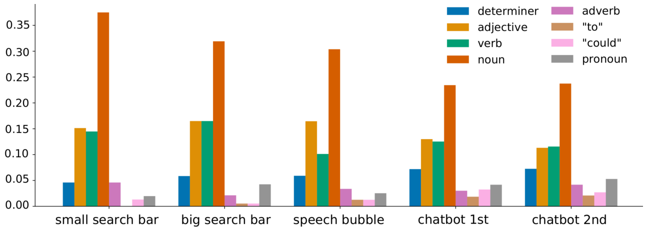


Figure 3: Distribution of most frequent parts of speech within each condition.

4.2 Natural Language

Besides the information content, we aim to investigate the usage of natural language – both concerning the interface design, as well as the information value level.

First, we investigate the sentence completeness by analysing whether participants formulated their query as a keyword search, using partial sentences, or using full sentences. In I4, 43% of queries are grammatically complete sentences, e.g.:

"I want a laptop that is designed by apple for business purposes. Between £700-£1000 and rose gold"

In I1, I2, and I3, only 18%, 12%, and 11% of the participants used complete sentences. Compared to I1, the usage of complete sentences increased significantly by 139% in I4. But also compared to I2 and I3, I4 brings about significantly more complete sentences. Keyword search shows the inverse trend, e.g.:

"windows laptop 8gb RAM"

Only 3% of the participants in I4 formulated their query in such a bullet point form, compared to 61% in I1 and 53% in I2. The results show a weakly positive correlation between sentence completeness and the number of key facts mentioned in a query (r = .20): The more complete the sentences were formulated, the more key facts were mentioned in the query.

To analyse the query formulation in more detail for natural language, Figure 3 presents the distribution of the eight most frequent grammatical types of words (parts of speech). While keyword searches should mainly consist of adjectives and nouns, grammatically complete sentences should contain a broader range of parts of speech. The results show that across conditions, the distribution and share of POS-tags changes.

Vagueness is an inherent characteristic of natural language. In our experiment, we found significantly more vague words mentioned in I4 after the follow-up question (M = 3.3, STD = 2.7) than in either of the non-chatbot conditions, as Figure 2c shows. Looking at the percentage of vague words in key facts, no differences between conditions can be found (p = .972). Thus, the increase in vague words in I4 can be attributed to an increase in query length; key facts are not described more vaguely in any condition. The sentence completeness, however, correlates moderately positively with query vagueness (r = .50). Vagueness was more often used in the context of natural language, e.g.:

> *"I'd like something* ***lightweight****,* ***easy*** *to use with a* ***long*** *battery life"*

Our results furthermore indicate that more domain knowledge reduces the number of vague words in a query (r = -.19).

4.3 Perception of Participants

After issuing the query, participants answered open questions about their experience with the search interface. The most striking observation was the repeated comparison with known search engines. 52% of the participants mentioned in some way that they are not used to natural language input for search engines:

> *"I wouldn't normally write in full sentences so it didn't feel natural."*

Participants described the placeholder text in the query as a major signal for natural language, since it was a partial grammatical sentence. Participants furthermore suggested to use choice questions ("A or B?") during a dialogue, react specifically to the initial query to show understanding, allowing voice input, and allowing to rank the mentioned attributes.

5 DISCUSSION

In this study, we found that the chatbot interface elicits longer queries with a significantly higher number of key facts, especially after posing a follow-up question. Furthermore, the diversity of attributes was greater for the speech bubble and chatbot interface, as compared to both search bar interfaces. These findings suggest that the interface design has a major influence on the amount of information a user reveals in the initial query (**RQ 1**). We also found that longer queries contain more key facts and that participants with greater domain knowledge write longer queries. However, there is no correlation between the number of key facts and domain knowledge. This means that other factors besides domain knowledge drive the formulation of long queries with higher numbers of key facts. One possible driving factor for more key facts could be natural language.

While vagueness is a characteristic of natural language, we did not find evidence that people are more inclined to use vague language in our dialogue-inspired conditions. It could be that participants did not use vague words for convenience, but rather due to a lack of domain knowledge. We see that when domain knowledge increases, vagueness decreases. The lack of knowledge is not compensated by the interface, which results in equally vague formulations across conditions. If participants want a "fast" laptop, but do not know what facet this attribute contains, they will not know it regardless of the interface. The results, however, show a clear trend towards partial or full grammatical sentences in the chatbot interface, and the more sentence-like a query, the more key facts are mentioned. Together with the observed change in usage of parts of speech, we conclude that the queries of the chatbot interfaces are formulated in a more natural way than the queries in the search bar conditions (**RQ 2**).

Overall, the results demonstrate two prospects. (1) It is possible to influence and steer the formulation of queries towards a more natural language. (2) It is possible to stimulate the user of a product search system to reveal more (and more diverse) information about the desired product already in the initial query, but even more so with a generic follow-up question – which is in line with the results reported in [2]. When stimulating users to reveal more information on the search target, search systems need to be equipped with appropriate functionality to process such long and diverse queries.

6 CONCLUSION

This study set out to investigate whether the interface design of a search engine can influence the amount of information revealed about a user's information need. We designed and executed an online user study (N=139) to investigate the query formulation for four search interfaces with varying conversational elements. Our key findings show that the interface design influences how and how much information users input into a search engine for the initial query. More precisely, conversation-like interfaces elicit more information and more natural language formulations at an initial stage of the information-seeking process than traditional search bar interfaces. This study reveals insights into the design of product search engines. For gaining a holistic image of the results across domains, follow-up research should investigate a broader variety of tasks to test the generalisability of the findings.

ACKNOWLEDGMENTS

This work was partly funded by the DFG, grant no. 388815326. We thank Alex Abas, Alexandra Bergen, Ali Khalife, Charlott Melcher, Annika Müller, Vanessa Sinanaj & Saskia Zarges for their work on the interface designs and the study.

REFERENCES

[1] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. 2013. Leading People to Longer Queries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3019–3022. https://doi.org/10.1145/2470654.2481418

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3331184.3331265

[3] Peter Bailey, Ryen W. White, Han Liu, and Giridhar Kumaran. 2010. Mining Historic Query Trails to Label Long and Rare Search Engine Queries. *ACM Trans. Web* 4, 4, Article 15 (Sept. 2010), 27 pages. https://doi.org/10.1145/1841909.1841912

[4] Evelyn Balfe and Barry Smyth. 2004. Improving Web Search through Collaborative Query Recommendation. In *Proceedings of the 16th European Conference on Artificial Intelligence* (Valencia, Spain) *(ECAI'04)*. IOS Press, NLD, 268–272.

[5] Catalin-Mihai Barbu, Guillermo Carbonell, and Jürgen Ziegler. 2019. The Influence of Trust Cues on the Trustworthiness of Online Reviews for Recommendations. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (Limassol, Cyprus) *(SAC '19)*. Association for Computing Machinery, New York, NY, USA, 1687–1689. https://doi.org/10.1145/3297280.3297603

[6] N. J. Belkin, C. Cool, D. Kelly, G. Kim, J. y. Kim, H. j. Lee, G. Muresan, M. c. Tang, and X. j. Yuan. 2002. Rutgers interactive track at TREC 2002. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002*. Addison Wesley.

[7] N. J. Belkin, 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. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 205–212. https://doi.org/10.1145/860435.860474

[8] Michael Bendersky and W. Bruce Croft. 2008. Discovering Key Concepts in Verbose Queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 491–498. https://doi.org/10.1145/1390334.1390419

[9] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The Vocabulary Problem in Human-System Communication. *Commun. ACM* 30, 11 (Nov. 1987), 964–971. https://doi.org/10.1145/32206.32212

[10] Djoerd Hiemstra, Claudia Hauff, and Leif Azzopardi. 2017. Exploring the Query Halo Effect in Site Search: Leading People to Longer Queries. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 981–984. https://doi.org/10.1145/3077136.3080696

[11] Yvonne Kammerer and Maja Bohnacker. 2012. Children's Web Search with Google: The Effectiveness of Natural Language Queries. In *Proceedings of the 11th International Conference on Interaction Design and Children* (Bremen, Germany) *(IDC '12)*. Association for Computing Machinery, New York, NY, USA, 184–187. https://doi.org/10.1145/2307096.2307121

[12] Diane Kelly and Xin Fu. 2007. Eliciting better information need descriptions from users of information search systems. *Information Processing & Management* 43, 1 (2007), 30 – 46. https://doi.org/10.1016/j.ipm.2006.03.006

[13] L. Palopoli, D. Rosaci, and D. Ursino. 2006. Agents' roles in B2C e-commerce. *AI Commun.* 19 (2006), 95–126.

[14] Andrea Papenmeier, Alfred Sliwa, Dagmar Kern, Daniel Hienert, Ahmet Aker, and Norbert Fuhr. 2020. 'A Modern Up-To-Date Laptop' - Vagueness in Natural Language Queries for Product Search. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 2077–2089. https://doi.org/10.1145/3357236.3395489

[15] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Stockholm, Sweden, 353–360. https://doi.org/10.18653/v1/W19-5941