Ah, Alright, Okay! Communicating Understanding in Conversational Product Search

Andrea Papenmeier
a.papenmeier@utwente.nl
University of Twente
Enschede, Netherlands

Elin Anna Topp
elin_anna.topp@cs.lth.se
Lund University
Lund, Sweden

ABSTRACT

When talking about products, people often express their needs in vague terms with vocabulary that does not necessarily overlap with product descriptions written by retailers. This poses a problem for chatbots in online shops, as the vagueness and vocabulary mismatch can lead to misunderstandings. In human-human communication, people intuitively build a common understanding throughout a conversation, e.g., via feedback loops. To inform the design of conversational product search systems, we investigated the effect of different feedback behaviors on users' perception of a chatbot's competence and conversational engagement. Our results show that rephrasing the user's input to express what was understood increases conversational engagement and gives the impression of a competent chatbot. Using a generic feedback acknowledgment (e.g., "right" or "okay"), however, does not increase engagement or perceived competence. Auto-feedback for conversational product search systems therefore needs to be designed with care.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Human computer interaction (HCI)**.

KEYWORDS

Conversational User Interface, Product Search, Chatbot, Feedback, Grounding

ACM Reference Format:
Andrea Papenmeier and Elin Anna Topp. 2023. Ah, Alright, Okay! Communicating Understanding in Conversational Product Search. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3571884.3604318

1 INTRODUCTION

Many people search for and buy products online. Online retailers offer a vast choice of products that customers can search through. However, while some searchers know exactly what they are looking for, others do not have a specific product in mind [22]. Some users have vague needs that are refined throughout the search [15]. Even if users know what they want, they often express their needs in vague, natural language [20] and use a different vocabulary than the product descriptions in the online shop [13]. Therefore, conversations in product search have a high risk of leading to misunderstandings.

In human-human conversations, conversation partners engage in *grounding* to mutually communicate their understanding of what was said [5, 16], leading to a common understanding and fewer misunderstandings. Such behavior can also be observed in sales dialogues: Experts often reflect their understanding by repeating what was said or communicating that they understood the customer with feedback such as "mhm" or "okay" [18].

Dialogue systems can be equipped with such feedback behavior. For example, repetition, summarization, and paraphrasing are means to express what was understood [1, 25]. Sophisticated grounding behavior that makes transparent what the system has understood from the user's input has been shown to increase conversational engagement in movie recommender systems [9].

However, it remains unclear how dialogue systems in product search can profit from auto-feedback, even though it is essential that the system correctly understands the user's needs. Therefore, we set out to explore how users perceive a product search chatbot that communicates its understanding via auto-feedback. Engaging in grounding and disclosing its inner state could make users perceive the conversation richer and more engaging. Another aspect could be that users are reassured that the chatbot is able to correctly process their input, which might increase the perceived competence of the dialogue system.

We conducted a user study to compare three chatbot behaviors with varying levels of feedback on the chatbot's understanding of prior user input. Participants were confronted with pre-recorded chatbot dialogues and reported their preception of the systems. In our study, simply acknowledging the user input with a generic message such as "right" or "okay" did not increase the perceived competence or conversational engagement. However, our findings show that a chatbot that discloses what it understood from the user's input is perceived as more competent and conversationally engaged than chatbots that do not engage in grounding behavior. With this study, we contribute empirical findings to inform the design of product search dialogue systems: When being employed in a use case with an increased need for grounding, such as customer-expert settings in e-commerce, chatbots should inform users about what information they extracted from the user's input rather than displaying a generic confirmation.

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3604318

2 RELATED WORK

Conversations are not just an exchange of information about the subject under discussion. Dialogue partners also communicate meta-information about the conversation, such as whether one understood what was said. To achieve a common understanding, participants of a conversation – human or computer – need to "coordinate their distinct knowledge states" [2], i.e., find a common ground. Grounding is an incremental and iterative feedback process [2]. With higher costs of failure, the steps between the feedback become smaller [2]. For example, if a person wants to buy an expensive product, it is important that the salesperson understands what the person wants. Petukhova and Manzoor [21] distinguish six levels of communicating understanding, e.g., recognition (nodding, "mhm"), interpretation (repeating and rephrasing), or evaluation (relating the understood information to one's own ideas and feelings). Focusing on individual acts within a conversation, Bunt [3] describes two types of feedback in a dialogue that express understanding of processing: "auto-feedback" that concerns the beliefs about one's own information processing, and "allo-feedback" which relates to beliefs about the other's processing of information. As such, auto- and allo-feedback types are part of grounding in dialogues, i.e., the act of building a common understanding [5, 16]. Dialogue systems need to incorporate these types of feedback for effective communication [3].

GoDiS [12] is such a dialogue system that takes a rudimentary approach to grounding – if the user does not correct the system, it means that the system correctly understood the user. In medical consulting, dialogue systems have been equipped with grounding skills to encourage users' self-disclosure and increase comfort [1, 25], e.g., summarizing or repeating what was understood. For a movie recommender, Grimes et al. [9] found that reflecting what was understood leads to higher conversational engagement compared to a system that only acknowledges user input. Besides usage of auto-feedback, Frummet et al. [8] also observed re-assurance behavior via repetitions in conversational search for cooking. Standardized protocols that include meta-information exchange for dialogue systems are scarce. Kiesel et al. [10] therefore proposed a framework for incorporating meta-information like auto-feedback into dialogues of conversational search.

In product search interactions, there is a great potential for grounding. On the one hand, people do not always have a clear vision of what they are searching for and use vague terms when talking about products [20]. Vagueness and technical terms still pose a problem to conversational agents [4], as the interpretation of vague expressions varies from person to person [17]. Grounding can help search systems in those cases to disambiguate user needs [24]. On the other hand, different people use different vocabulary. In online shopping, for example, customers use a different vocabulary than retailers, but also retailers among themselves do not have a consistent vocabulary [13]. In human-human conversations, resolving such vocabulary mismatches are part of grounding [2]. By observing product search dialogues between experts and customers, Papenmeier et al. [18] found that especially experts make use of auto-feedback such as "mhm", "okay", or literal repetitions.

3 METHOD

This study set out to assess how auto-feedback influences users' perception of a chatbot in product search. Auto-feedback can be expressed in form of a generic acknowledgement, e.g., "mhm" or "okay", or be an informative statement, e.g., repetition, summarization, or paraphrase [1, 9, 18, 21, 25]. In product search, experts often use repetitions and acknowledgments toward customers [18]. We compared the following three behaviors:

NONE: The system does not react to the user's input (see Figure 1a).

ACK: The system acknowledges that it has understood the user's input with a generic feedback of understanding such as "Right" or "Okay" (see Figure 1b).

PARA: The system communicates what it has understood from the user's input with an informative feedback of understanding by paraphrasing (see Figure 1c).

We employed a quantitative research approach to understand the effect of auto-feedback in product search conversations. The study uses a within-subject design with pairwise comparison of auto-feedback behaviors, i.e., comparing *NONE* to *ACK*, *NONE* to *PARA*, and *ACK* to *PARA*. That is, each participant assessed two behaviors to reduce the mental load and complexity for participants and avoid fatigue effects when being confronted with very similar dialogues. Participants saw the behaviors in randomized order to counteract order effects.

The user study received ethical clearance from the first author's institute's ethics committee.

3.1 Use Case and Scenario

To investigate auto-feedback in product search, we chose laptop search as users have vague needs in this product category (see [20]) that raise the need for grounding. Moreover, we expected a vocabulary mismatch for technical terms between customers and retailers in this product category. Users might be unsure if their input is correctly phrased and would profit from grounding.

We generated three product search dialogues with chatbots that show *NONE*, *ACK*, or *PARA* behavior as a reaction to user input. The dialogues included a welcoming message, five questions on desired laptop attributes (see Figure 1 for excerpts), and a final message saying that the chatbot will now search for suitable laptops. The user inputs in the dialogues were generated by a native English speaker from the UK who was unfamiliar with the research project.

In the user study, we introduced the dialogues with a scenario describing the dialogue context: *"Imagine your friend's laptop broke down. You observe how your friend searches for a new laptop on the internet. Your friend tests two different websites that offer a dialogue system to search for a new laptop."* The scenario introduced the search as done by a third party (the friend), allowing us to use screenshots of a dialogue instead of an interactive system.

3.2 Study Procedure

Respondents completed the user study online, on their own devices. Participation was restricted to computers and laptops to ensure a correct display of the dialogues. After giving informed consent, participants indicated their demographic background (gender, age) and their affinity for technology interaction. Subsequently, the

scenario was presented as shown above. On the next page, the first dialogue screenshot was displayed. Participants then rated the perceived competence of the chatbot and the conversational engagement below the screenshot. The next page followed the same structure, with the second dialogue displayed. After rating both dialogues, participants were asked which system they preferred and why before receiving the debriefing text with additional information on the research goal. The online survey, including all dialogue screenshots, are available online[1].

3.3 Measures and Analysis

We measured users' perception of the auto-feedback behaviors with two dependent variables:

Conversational Engagement: Based on [9], we measure conversational engagement with four items (statements on politeness, conversational skills, engagement, and communication like a human) that participants rate on a 7-point scale from "strongly disagree" to "strongly agree". The conversational engagement score is calculated as the average score of the four items.

Perceived Competence: We define the perceived competence of the chatbot as the perceived likelihood that it will perform its task as expected. Participants rated "How likely will the system retrieve a laptop that matches your friend's needs?" on a 6-point scale from "very unlikely" to "very likely".

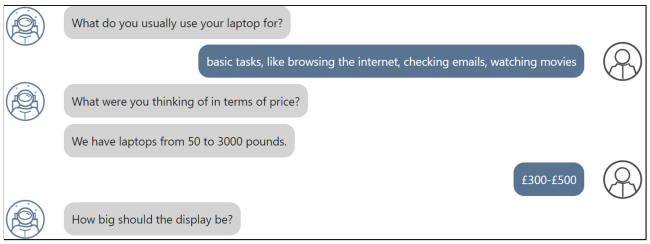
To gain additional insights into how participants perceived the different auto-feedback behaviors, we asked them to indicate their preference for a system and explain their choice in an open question.

We performed Wilcoxon's signed rank test for pairwise comparison of dependent samples to compare auto-feedback behaviors as the data was not normally distributed. We adjusted the p-values with the Bonferroni method to correct for repeated tests. We report effect sizes with Cohen's d.

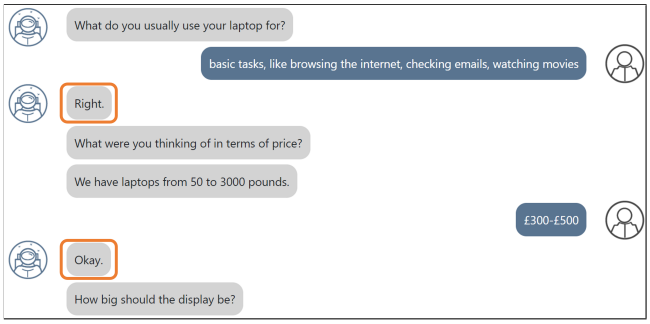
3.4 Participants

We recruited 105 participants (English native speakers, no literacy difficulties, UK residents) on Prolific [2]. Based on participants' answers to the open questions, none of the responses had to be excluded for low effort or low quality. The participants were randomly assigned two distinct dialogues, resulting in three groups: 35 participants comparing *NONE* to *ACK*, 35 comparing *NONE* to *PARA*, and 35 comparing *ACK* to *PARA*. On average, the respondents (49 female, 55 male, 1 non-binary, 0 unreported) were 42.7 years old (SD = 14.5). The ANOVA did not show a significant main effect regarding differences in age ($F(1,102) = 0.197$, $p = 0.6604$) or gender distribution ($F(1,102) = 0.314$, $p = 0.7316$). Personal affinity for interacting with technology might impact how a participant perceives new technology. We therefore tested for differences in affinity for technology interaction (nine items, see [7]) but did not find a significant main effect ($F(2,102) = 0.004$, $p = 0.9519$) among the three groups.

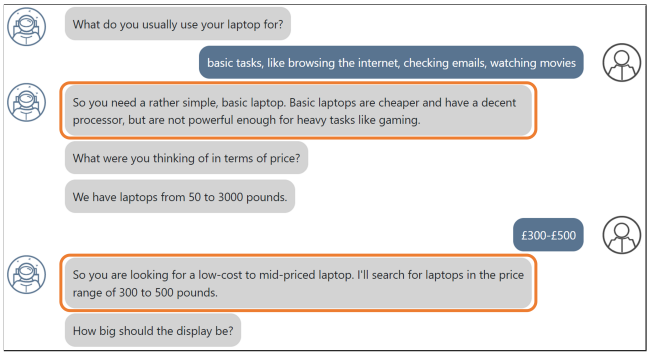
[1] https://git.gesis.org/papenmaa/cui23_communicatingunderstanding
[2] https://www.prolific.co



(a) *NONE*



(b) *ACK*



(c) *PARA*

Figure 1: Excerpts of the product search chatbot with (a) no feedback of understanding, (b) generic feedback by acknowledging, and (c) informative feedback of understanding by paraphrasing. The utterances expressing understanding are circled in orange.

3.5 Limitations

The study setup is subject to several limitations. The study relies on a hypothetical scenario with screenshots of dialogues rather than live interaction. Allowing users to experience the different auto-feedback behaviors on their own inputs could impact the results. However, an interactive system introduces several biases, e.g., through technical malfunctions or varying dialogue lengths. Approaches using screenshots have been reported in literature for evaluating interactive systems, e.g., conversational systems [6, 23] or retrieval systems [19]. This study uses the same pre-recorded dialogue across all participants. Although this avoids varying dialogue lengths, it reduces the generalizability of the findings. To counteract this effect, we chose a dialogue recording that represents generic, average needs of the population. Due to these limitations,

the findings of this study should serve as a starting point for further explorations rather than provide a comprehensive investigation of auto-positive feedback types.

4 RESULTS

Concerning the level of **conversational engagement** (see Figure 2a), dialogues in which the chatbot communicated what it has understood by paraphrasing (*PARA*) were perceived to be significantly more engaging (moderate effect) than only acknowledging the user input (*ACK*) or not giving auto-feedback to the user input (*NONE*). Most comments of participants concerned the conversational skills of the system. They described *PARA* as more in-depth (11 of 35 comparing *PARA* to *ACK*, 8 of 35 comparing *PARA* to *NONE*). Participants explained the advantage of receiving a more detailed feedback: "the decision feels more informed" and "when you[']re spending a lot of money on a computer you want engagement". Another suggested that it "could be really helpful for those with less knowledge". However, acknowledging the user input with generic auto-feedback words such as "okay" (*ACK*) did not significantly increase the conversational engagement compared to not providing auto-feedback (*NONE*). Nevertheless, participants noticed the difference between the dialogues: Several participants who had compared *NONE* to *ACK* reported that *NONE* was more "to the point", i.e., used less superfluous words that slowed down the conversation (11 of 35). Others explained that *ACK* was more polite, seemed more human-like, friendlier, and more willing to help (10 of 35).

To investigate the effect of different auto-feedback behaviors on **perceived competence**, participants rated how likely the system will retrieve results matching their friend's needs. In direct comparison, disclosing information on what was understood (*PARA*) yielded significantly higher perceived competence ratings than both the *ACK* and the *NONE* auto-feedback behavior (small to moderate effect sizes). Some participants commented on how the paraphrasing auto-feedback related to their feeling of being understood: "It is reassuring that the robot [...] demonstrates understanding of the customers needs" and "I don't really need them to feel like they're talking, I just want to know that they actually understand what I'm saying". One participant related the auto-feedback directly to the chances of seeing relevant products after the dialogue: "[B]y giving more detailed feedback[,] it also gives the customer more information on the kind of laptops they will be presented with". In contrast, there was no significant difference between *ACK* and *NONE* (see Figure 2b). Out of 35 participants that had seen *ACK* and *NONE*, only one addressed how acknowledging expresses information about being able to handle the user's input: "saying "right" or "okay" [...] makes you feel that it is more likely to be able to help you find what your looking for".

5 DISCUSSION AND CONCLUSION

In our experiment, *PARA* received significantly higher competence and conversational engagement scores than the other two auto-feedback behaviors. With competence being an aspect of credibility [11], this auto-feedback behavior might also positively affect the overall credibility of the chatbot. Especially in e-commerce, expressing competence is a desired characteristic of chatbots as it increases not only trust in the online shop, but also the purchase intentions [14]. However, as this behavior requires advanced natural language processing, it is the most difficult to implement.

Practitioners might be inclined to implement the acknowledging behavior (*ACK*) to reduce implementation effort. However, we did not see positive effects on conversational engagement or competence between *ACK* and *NONE*. Contrarily, some users might even prefer not to have feedback at all because generic, uninformative feedback can feel redundant. In our use case, acknowledging with a generic, uninformative statement could not serve as an alternative to the the more implementation-heavy rephrasing behavior. In human-human conversations, literature has frequently found acknowledgments such as "mhm" or "okay" being used between interaction partners [8, 18]. However, our findings indicate that this might not extend to human-machine conversations. This discrepancy might arise because users of conversational systems do not always know the extent of the system's processing and sense-making abilities. That is, a system might have to show users explicit proof that it has received and correctly interpreted the input.

The findings provide first insights into feedback mechanisms for grounding in product search. Grounding typically consists of several feedback loops [2]. Future research should extend the findings from our static hypothetical conversations and investigate interactive settings with multiple interactive feedback cycles over time. Moreover, besides rephrasing and acknowledging, other feedback types such as summarization or repetition [1, 18, 25] are possible. Future studies could investigate those additional feedback types or mixtures of types, e.g., paraphrasing to prove the system's capabilities, followed by the shorter acknowledging behavior.

REFERENCES

[1] Fahad Almusharraf, Jonathan Rose, and Peter Selby. 2020. Engaging Unmotivated Smokers to Move Toward Quitting: Design of Motivational Interviewing–Based Chatbot Through Iterative Interactions. *Journal of Medical Internet Research* 22, 11 (2020), 1–14. https://doi.org/10.2196/20251

[2] Susan E Brennan. 1998. The grounding problem in conversations with and through computers. In *Social and cognitive approaches to interpersonal communication* (1 ed.). Lawrence Erlbaum, 201–225.

[3] Harry Bunt. 2012. The semantics of feedback. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. SEMDIAL, Paris, France. http://semdial.org/anthology/Z12-Bunt_semdial_0016.pdf

[4] Julia Cambre, Ying Liu, Rebecca E. Taylor, and Chinmay Kulkarni. 2019. Vitro: Designing a Voice Assistant for the Scientific Lab Workplace. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 1531–1542. https://doi.org/10.1145/3322276.3322298

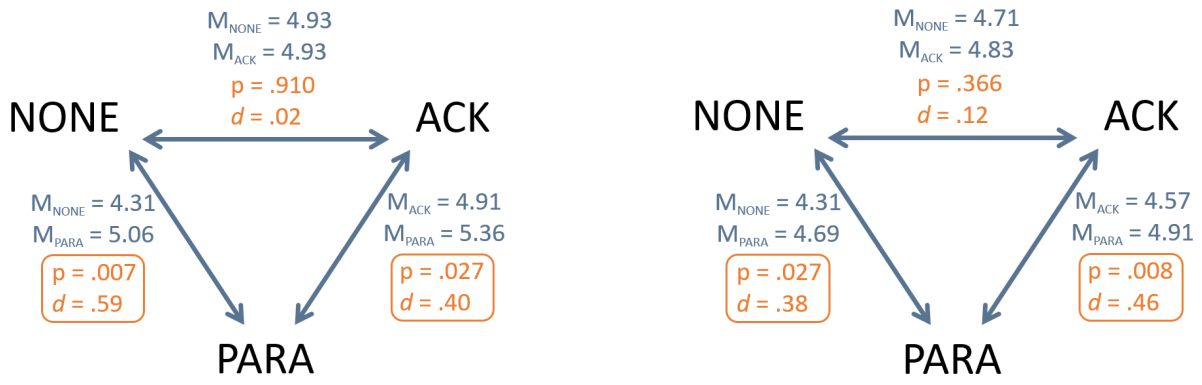
[5] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*. American Psychological Association, 127–149. https://doi.org/10.1037/10096-006

[6] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. https://doi.org/10.1145/3543829.3543831

[7] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[8] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What Can I Cook with These Ingredients?" - Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems* 40, 4, Article 81 (2022), 32 pages. https://doi.org/10.1145/3498330

[9] G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems* 144 (2021), 113515. https://doi.org/10.1016/j.dss.2021.113515



(a) Conversational Engagement **(b) Perceived Competence**

Figure 2: Visualization of the (a) conversational engagement and (b) perceived competence scores. We report mean values, p-values of Wilcoxon's signed rank test (displayed after Bonferroni correction), and effect sizes. Significant effects are circled in orange.

[10] Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. 2021. Meta-Information in Conversational Search. *ACM Trans. Inf. Syst.* 39, 4, Article 50 (aug 2021), 44 pages. https://doi.org/10.1145/3468868

[11] Philipp Kulms and Stefan Kopp. 2016. The effect of embodiment and competence on trust and cooperation in human–agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 75–84. https://doi.org/10.1007/978-3-319-47665-0_7

[12] Staffan Larsson, Peter Ljunglöf, Robin Cooper, Elisabet Engdahl, and Stina Ericsson. 2000. GoDiS: An Accommodating Dialogue System. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational Systems - Volume 3* (Seattle, Washington) *(ANLP/NAACL-ConvSyst '00)*. Association for Computational Linguistics, USA, 7–10. https://doi.org/10.3115/1117562.1117564

[13] Aarno Lehtola, Johannes Heinecke, and Catherine Bounsaythip. 2003. Intelligent human language query processing in MKBEEM. In *Proceedings of the International conference on Universal Access in Human-Computer Interaction (UAHCI'03)*. 22–27.

[14] Tze Wei Liew, Su-Mae Tan, Jessica Tee, and Gerald Guan Gan Goh. 2021. The effects of designing conversational commerce chatbots with expertise cues. In *14th International conference on human system interaction (HSI '21)*. IEEE, 1–6. https://doi.org/10.1109/HSI52170.2021.9538741

[15] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[16] Andrew Monk. 2003. Common ground in electronically mediated communication: Clark's theory of language use. *HCI models, theories, and frameworks: Toward a multidisciplinary science* (2003), 265–289.

[17] Daniel R. Montello, Michael F. Goodchild, Jonathon Gottsegen, and Peter Fohl. 2003. Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation* 3, 2-3 (2003), 185–204. https://doi.org/10.1080/13875868.2003.9683761

[18] Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "Mhm..." – Conversational Strategies For Product Search Assistants. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 36–46. https://doi.org/10.1145/3498366.3505809

[19] Andrea Papenmeier, Daniel Hienert, Firas Sabbah, Norbert Fuhr, and Dagmar Kern. 2022. UNDR: User-Needs-Driven Ranking of Products in E-Commerce. In *Workshop On eCommerce (SIGIR eCom)*. 1–6.

[20] 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*. Association for Computing Machinery, New York, NY, USA, 2077–2089. https://doi.org/10.1145/3357236.3395489

[21] Volha Petukhova and Hafiza Erum Manzoor. 2021. Towards the ISO 24617-2-compliant Typology of Metacognitive Events. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Association for Computational Linguistics, Groningen, The Netherlands (online), 14–19. https://aclanthology.org/2021.isa-1.2

[22] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A Taxonomy of Queries for E-Commerce Search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1245–1248. https://doi.org/10.1145/3209978.3210152

[23] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300768

[24] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management* 57, 2 (2020), 102162. https://doi.org/10.1016/j.ipm.2019.102162

[25] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376131