An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues

Svitlana Vakulenko
University of Amsterdam
Amsterdam, The Netherlands
s.vakulenko@uva.nl

Evangelos Kanoulas
University of Amsterdam
Amsterdam, The Netherlands
e.kanoulas@uva.nl

Maarten de Rijke
University of Amsterdam & Ahold Delhaize
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

The ability to engage in mixed-initiative interaction is one of the core requirements for a conversational search system. How to achieve this is poorly understood. We propose a set of unsupervised metrics, termed *ConversationShape*, that highlights the role each of the conversation participants plays by comparing the distribution of vocabulary and utterance types. Using ConversationShape as a lens, we take a closer look at several conversational search datasets and compare them with other dialogue datasets to better understand the types of dialogue interaction they represent, either driven by the information seeker or the assistant. We discover that deviations from the ConversationShape of a human-human dialogue of the same type is predictive of the quality of a human-machine dialogue.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

Conversational search, mixed initiative, dialogue

ACM Reference Format:
Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2020. An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3397271.3401297

1 INTRODUCTION

While the idea of conversational search has been around for several decades [see, e.g., 3], the idea has recently attracted considerable attention [21]. Conversational user interfaces are believed to facilitate more efficient information access than traditional interfaces. A conversation, in this case, is a collaborative process that allows an information seeker to satisfy an information need. One of the key features of a conversational interaction is the potential for *mixed initiative*, where "the system and user both can take initiative as appropriate" [21]. What is an appropriate action on the part of the system? How initiative can be measured? In this paper we report on the first attempts to analyse and evaluate the degree of initiative and collaboration between the conversation participants given only conversation transcripts as input.

Conversational search tasks proposed to date mostly reduce dialogues to a sequence of question-answer pairs [2, 8, 20]. In the datasets used for the question-answering tasks, the structure of interactions is fixed in advance: either the user takes the initiative and the system follows-up with an answer or vice versa, which makes them unsuitable for studying how the initiative is transferred between roles. Discussion threads from on-line Q&A forums are a popular data source for developing conversational search tasks [18, 19]. While on-line forums are a valuable resource for studying real-world interaction patterns, they exhibit a type of asynchronous information exchange, which, as we show in our analysis, is very different from synchronous dialogue interactions.

Conversational systems are often grouped into question answering, task-oriented and chit-chat [12]. It is important to note, that this classification schema is mainly based on the difference in approaches to building such conversational systems rather than the difference in dialogues they produce. In this paper, we focus on analysing and measuring differences between dialogue types, and report on the resulting dimensions and a new dialogue classification scheme that emerges from our analysis. We show that human-to-human dialogues collected for conversational search tasks bear structural similarities to task-oriented and chit-chat dialogues.

Recent evaluation studies of chit-chat dialogue models show that conversational systems tend to seize control of the dialogue by asking too many questions and ignoring the user's initiative [4, 9]. Standard evaluation metrics do not capture this dimension of a dialogue interaction and, therefore, fail to predict user engagement [9]. The most popular metric for dialogue evaluation is relevance of the response, which is usually measured with respect to a ground-truth response; it is comparable to answer accuracy if the response is an answer. Our work is complementary to this research. We introduce a novel evaluation framework based on a set of unsupervised features. The framework is designed to capture the quality of dialogue interactions in terms of balancing initiative, when appropriate, and measuring collaboration between the dialogue participants.

Our evaluation framework is based on several independent lexical features that capture initiative and collaboration in dialogues. Simple automated measures based on discourse features, such as lexical and syntactic diversity, were previously adopted to reduce repetitive generic responses and to estimate question complexity [14, 23]. We use an unsupervised approach, similar to the ones applied in language style matching [13] and in measuring quality of generated narratives [22], to a dialogue setting. A key characteristic of a dialogue is

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401297

that it is a type of narrative with utterances generated by multiple dialogue participants. Therefore, we estimate lexical features separately for each participant so as to be able to compare their contributions and, thereby, deduce the roles they play in the conversation.

The work most similar to ours is by Walker and Whittaker [25], who studied lexical cues, such as the use of anaphora and different utterance types, as a mechanism for switching control between dialogue participants. Our approach to dialogue representation is unsupervised and domain-independent, which allows us to scale an analysis that was previously performed only on a handful of dialogues to thousands of publicly available dialogue transcripts.

Our main contributions can be summarised as follows: (1) We examine structural patterns of initiative and collaboration across ten datasets with more than 97k dialogues. Ours is the first study that automatically identifies these dimensions within large and diverse dialogue corpora, drawing parallels between dialogue tasks that originate within different research communities. (2) The initiative and collaboration patterns we identified correlate with human judgements of dialogue quality. Allocation of control, where control is defined as managing the direction of flow in a conversation, is at the core of mixed-initiative dialogue systems that are designed to enhance human-machine collaboration [7]. Dialogue systems should be able to recognize the user's cues for initiative switch so as to provide appropriate responses [6]. Detecting initiative is also important to characterize the quality of the interaction [7]. Our work contributes insights that inform the design of evaluation and optimisation approaches capable of recognising initiative distribution in dialogue.

2 CONVERSATIONSHAPE

ConversationShape is a dialogue representation approach that focuses on the structural properties of a dialogue. We consider a dialogue to be a sequence of utterances exchanged between several participants. All dialogues in our experiments have two participants. However, our approach can be also applied to multi-party conversations. Information-seeking dialogues are often characterised by asymmetry of the roles that participants play in the conversation: one usually assumes the role of an assistant (A), whose function is to be automated by a conversational search system; another dialogue participant is an information seeker (S), who is using the service of the assistant to obtain information. To model mixed initiative in dialogues we use four metrics that are calculated separately for each of the dialogue participants: (1) question; (2) information; (3) repetition; and (4) flow.

Question is an explicit attempt at controlling the direction of a conversation flow, since a posed question sets an expectation for another participant to produce a relevant answer. We trained a supervised classifier on the NPS Chat Corpus [11] to recognize questions and other utterance types. The NPS Chat Corpus contains 7.9K utterances from online chat rooms, annotated with 14 utterance types ('Statement', 'Emotion', 'Greet', 'Bye', 'Accept', 'Reject', 'whQuestion', 'ynQuestion', 'yAnswer', 'nAnswer', 'Emphasis', 'Continuer', 'Clarify', 'Other'). Our classification model was initialised from a pre-trained RoBERTa [17] (base model) and further fine-tuned for the utterance type prediction task achieving F1 of 0.81 on the held-out test set.

The remaining metrics describe patterns of collaboration and control over the topic of a conversation. To explain them, we need to introduce the concept of dialogue vocabulary first. The *dialogue vocabulary* consists of all unique words (or subword tokens) that occur in the same dialogue transcript. We are especially interested in the words that occur frequently (more than once) within the same dialogue, since the repetition patterns are likely to signal their importance to the topic of a dialogue.

Information reflects the contribution that a participant made to the topic of a conversation. We estimate information as a count of frequent tokens that were first coined by a conversation participant.

Repetition indicates a follow-up on the topic of the conversation. To analyse the emergence of a shared vocabulary we trace vocabulary reuse patterns between the conversation participants. We estimate repetition as the number of tokens that were first introduced by one conversation participant and subsequently repeated by another conversation participant. We consider repetition as a type of relevance feedback available within the conversation, assuming the act of repetition to be an endorsement of the importance of the token to the topic of a conversation by increasing the token frequency. Another way to reference previous tokens implicitly is to use anaphora. Therefore, we add the count of anaphora to the count of repetitions. We use a short list of English anaphora from the analysis framework proposed by Walker and Whittaker: 'it', 'they', 'them', 'their', 'she', 'he', 'her', 'him', 'his', 'this', 'that'. We also experimented with an off-the-shelf co-reference resolution model instead but were not satisfied with the results.

Flow is the difference between *Repetition* and *Information*, which reflects on the role of a participant in maintaining the coherence of the conversation by referencing previous statements or driving the conversation forward by introducing new information.

For every conversation we compute the values for each of the conversation participants separately: $Concept_A$ and $Concept_S$, where *Concept* denotes one of the four metrics that we have just introduced. To be able to compare between conversations of different length we also normalise the scores by the number of utterances in a conversation. Then we use the average and the difference between the two metrics to characterise the type of dialogues in a dataset. The average shows the magnitude for each of the metrics, e.g. the average number of questions per conversation:

$$Concept = \frac{Concept_A + Concept_S}{2}. \tag{1}$$

The difference allows to compare the distribution (balance) between the dialogue participants, e.g. who asks more questions in a conversation. We use the formula similar to the one used for the writing style matching [13]:

$$\Delta Concept = \frac{Concept_A - Concept_S}{Concept_A + Concept_S}. \tag{2}$$

It not only indicates the difference in metrics between the roles but also its direction: negative values indicate dominance by the Seeker and positive – by the Assistant.

Example 2.1. Let us consider a snippet from the dialogue transcript in the Redial dataset [15] to illustrate our approach to measuring different aspects of initiative and collaboration in dialogue:

(A) Hey! What kind of *movies* do you like to watch?
(S) I'm really big on indie romance and dramas
(A) Ok what's your favorite *movie*?
(A) Staying with *that* genre, have you seen @88487 or @104253
(A) Those are two really good ones

(S) When I was a kid I liked *horror* like @181097

(A) @Misery is really creepy but really good. I only recently got into *horror*.

Assistant (A) clearly dominates the conversation asking all the questions ($Question_A = 2/7 = 0.29; Question_S = 0; Question = 0.15; \Delta Question = 1$). A introduced *movies* as the topic of the conversation but subsequently followed up on the topic directions introduced by S: *that* genre and *horror* ($Information_A = Information_S = 1/7 = 0.14; Information = 0.14; \Delta Information = 0$; $Repetition_A = 2/7 = 0.29$; $Repetition_S = 0$; $Repetition = 0.29/2 = 0.15$; $\Delta Repetition = 1$; $Flow_A = Repetition_A - Information_A = 0.29 - 0.14 = 0.15$; $Flow_S = -0.14$). We use this approach in the next section to automatically distinguish between dialogues of different types.

3 DATASETS

Our analysis spans across 10 publicly available dialogue datasets, which were designed for various dialogue tasks. Numbers in brackets indicate the number of dialogues in each of the datasets.

- **CCPE** (502) – conversational preference elicitation [20].
- **SCS** (37) – spoken conversational search [24].
- **MSDialog** (35.5K) – discussion threads from a support forum [19].
- **MultiWOZ** (10.4K) – multi-domain task-oriented dialogues [5].
- **ReDial** (10K) – conversational movie recommendation [15].
- **WoW** (22.3K) – chatting over topics from Wikipedia [10].
- **DailyDialog** (11K) – sample dialogues for English learners [16].
- **Meena** (91), **Mitsuku** (100), **Human** (95) – human-machine and human-human open-domain dialogues [1].
- **ConvAI2** (3.5K), **Control-M** (3.2K), **Control-H** (102) – human-machine and human-human persona-grounded dialogues [9, 23].

4 RESULTS

Table 1 shows the average *ConversationShape* for each of the dialogue sets from the previous section. This representation allows to compare the sets and identify different dialogue types, e.g., Figure 1 shows the clusters that emerge based on the similarities in Question and Information distributions.

Assistant-driven dialogues. From Table 1 we see that in CCPE the Assistant leads the conversation by asking the questions and the Seeker follows up by answering them (negative $\Delta Repetition$). MultiWOZ and MSDialog also have the majority of questions posed by the Assistant but those questions follow-up on the questions and answers provided by the Seeker (positive $\Delta Repetition$). In ReDial the Assistant drives the conversation by providing information and asking questions, while the Seeker follows up (negative $\Delta Repetition$).

Seeker-driven dialogues. SCS and WoW are similar to each other: for both the Seeker is mainly asking questions and the Assistant is providing information. However, the Seeker follows-up on the topics introduced by the Assistant in WoW (negative $\Delta Repetition$), while in SCS the Assistant follows the Seeker. Chit-chat dialogues (Human and Control-H) appear closer to the origin showing that the initiative is more balanced between the participants in this dialogue type. Whereas, in the DailyDialog dataset the initiative is skewed towards the initiator of the conversation, who is more likely to ask questions and set the conversation topic.

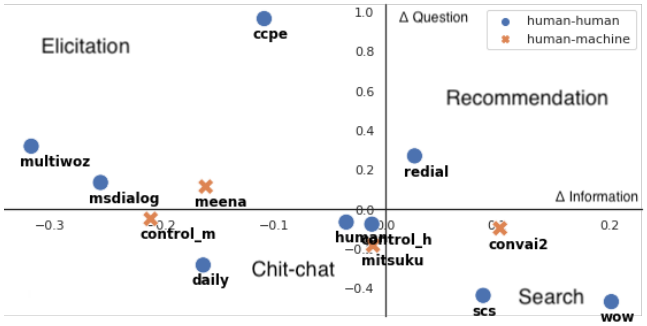

Figure 1: Dialogue types: datasets above the y-axis contain dialogues where questions are mostly posed by the Assistant (Assistant-driven dialogues), below – by the Seeker (Seeker-driven dialogues); in the dialogues on the right of the x-axis the conversation topic is mainly contributed to by the Assistant, on the left – by the Seeker.

Model diagnostics. ConversationShape can help to evaluate dialogue models and understand the type of deviant behavior the model exhibits. These experiments were performed on the subsets of the Control-M dataset that correspond to the transcripts produced by different dialogue models. In total, there are 28 models and we produce a ConversationShape profile for each of them. Then, we measure cross-entropy between the models' distributions and the distribution calculated for the subset of human-human dialogues (Control-H). Finally, we compare our results with the human evaluation results reported in the original paper [23]. The model with the lowest cross-entropy (0.01) to the human-human distributions was also the model that was preferred by human judges with respect to Interestingness and characterised by better flow and more information sharing. Moreover, ConversationShape also allows to interpret the type of deviation the model exhibit by observing each of the dimensions separately.

We plot all dialogue models from the Control-M set separately in Figure 2. Our metrics correctly identify outliers that either ask too many questions (optimised for inquisitiveness, *interviewer*), repeat too much (optimised for responsiveness, *parrot*) or do not follow up (optimised for diversity or negative responsiveness, *talker*). We could not achieve the same results when comparing the transcripts of Meena and Mitsuku dialogues [1]. The question distribution shows that Meena and Mitsuku dialogues are structurally very different from each other and from typical human chit-chat distribution. Mitsuku is being interrogated while Meena takes over the initiative by asking questions.

5 CONCLUSION

In this paper, we introduced the ConversationShape framework, which provides a set of simple but effective unsupervised metrics designed to measure initiative and flow of a conversation. Our analysis uncovers relations between different dialogue types and suggests a set of dimensions that are appropriate to consider when developing and evaluating conversational systems, or collecting new dialogue datasets. Our *Repetition* metric, which estimates follow-ups on a conversation topic is rather crude since it considers only lexical matches and anaphors. Though we show that it suffices for a high-level analysis of dataset distributions, predicting quality of the

Table 1: ConversationShapes of the popular dialogue datasets listing average and difference in Question, Information and Repetition between the conversation participants. Bold font highlights the highest values for each of the metrics. Grey marker indicates negative values, where the averages are skewed towards the Seeker.

Dataset	*Question*	Δ*Question*	*Information*	Δ*Information*	*Repetition*	Δ*Repetition*	*Flow*$_A$	*Flow*$_S$
CCPE [20]	0.21 (0.03)	**0.97** (0.10)	0.23 (0.10)	-0.11 (0.37)	0.29 (0.09)	-0.27 (0.30)	0.02 (0.12)	0.10 (0.16)
MultiWOZ [5]	0.22 (0.07)	0.32 (0.40)	0.39 (0.14)	**-0.32** (0.39)	0.44 (0.15)	0.32 (0.35)	0.31 (0.31)	-0.22 (0.28)
ReDial [15]	0.10 (0.05)	0.27 (0.56)	0.13 (0.08)	0.03 (0.58)	0.21 (0.09)	-0.20 (0.41)	0.04 (0.15)	0.12 (0.15)
MSDialog [19]	0.08 (0.09)	0.14 (0.68)	**1.13** (0.71)	-0.25 (0.53)	**1.02** (0.55)	**0.46** (0.42)	**0.64** (1.13)	**-0.85** (1.07)
WoW [10]	0.14 (0.08)	-0.47 (0.66)	0.37 (0.15)	0.20 (0.50)	0.52 (0.18)	-0.03 (0.37)	0.05 (0.38)	0.24 (0.35)
SCS [24]	0.14 (0.06)	-0.44 (0.39)	0.40 (0.28)	0.09 (0.39)	0.45 (0.21)	0.21 (0.35)	0.12 (0.40)	-0.02 (0.37)
DailyDialog [16]	0.17 (0.09)	-0.28 (0.69)	0.16 (0.15)	-0.16 (0.62)	0.30 (0.19)	0.08 (0.53)	0.20 (0.27)	0.08 (0.27)
Control-H [23]	0.20 (0.08)	-0.08 (0.46)	0.18 (0.10)	-0.01 (0.58)	0.22 (0.11)	0.10 (0.60)	0.05 (0.18)	0.02 (0.21)
Human [1]	0.21 (0.06)	-0.07 (0.35)	0.30 (0.13)	-0.04 (0.40)	0.36 (0.14)	0.02 (0.33)	0.08 (0.22)	0.03 (0.20)
Meena [1]	0.20 (0.06)	0.11 (0.47)	0.14 (0.06)	-0.16 (0.50)	0.17 (0.08)	0.21 (0.42)	0.08 (0.13)	-0.02 (0.11)
Mitsuku [1]	0.20 (0.06)	-0.18 (0.47)	0.15 (0.10)	-0.01 (0.53)	0.22 (0.10)	0.20 (0.39)	0.11 (0.19)	0.03 (0.16)
Control-M [23]	**0.25** (0.07)	-0.05 (0.43)	0.14 (0.09)	-0.21 (0.65)	0.16 (0.10)	0.06 (0.61)	0.07 (0.19)	-0.03 (0.20)
ConvAI2 [9]	0.18 (0.14)	-0.10 (0.63)	0.09 (0.10)	0.10 (0.50)	0.09 (0.11)	0.08 (0.50)	0.01 (0.19)	-0.00 (0.15)

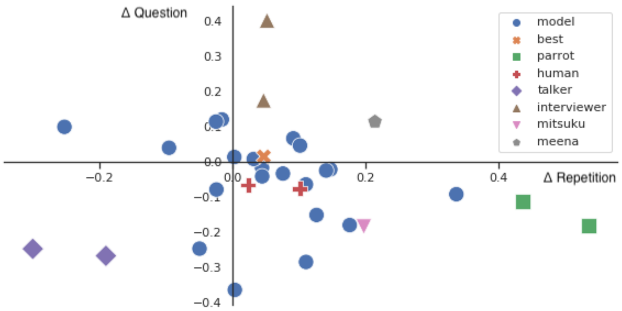

Figure 2: Model diagnostics with all models of the Control-M dataset unrolled. More successful dialogue models tend to ask more questions than in a typical human chit-chat (best and meena). This strategy adaptation does not imply, however, that these models are equally good at following the initiative and answering human questions.

individual dialogues requires a more fine-grained inspection. Future work should focus on developing an extension that can also account for semantic similarity between tokens. The next step will be to incorporate these metrics into an optimisation criteria of a learning algorithm that can supply the model with an appropriate perspective on the flow of a conversation and give an explicit incentive to control for an appropriate balance, which, as we showed, depends on the type of the conversation.

ACKNOWLEDGEMENTS

This research was supported by NWO (016.Vidi.189.039 and 314-99-301), H2020 (814961), and Google Faculty Research Award. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, et al. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Mohammad Aliannejadi, Hamed Zamani, et al. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR*. 475–484.

[3] Nicholas J. Belkin, Colleen Cool, et al. 1995. Cases, Scripts, and Information Seeking Strategies: On the Design of Interactive Information Retrieval Systems. *Expert Systems with Applications* 9 (1995), 379–395.

[4] Kevin K. Bowden, JiaQi Wu, et al. 2019. Entertaining and Opinionated But too Controlling: A Large-scale User Study of an Open Domain Alexa Prize System. In *CUI 2019*. 24:1–24:10.

[5] Pawel Budzianowski, Tsung-Hsien Wen, et al. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*. 5016–5026.

[6] Jennifer Chu-Carroll and Michael K. Brown. 1998. An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions. *User Modeling and User-Adapted Interaction* 8, 3-4 (1998), 215–254.

[7] Robin Cohen, Coralee Allaby, et al. 1998. What is Initiative? *User Modeling and User-Adapted Interaction* 8, 3 (1998), 171–214.

[8] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. TREC.

[9] Emily Dinan, Varvara Logacheva, et al. 2020. The Second Conversational Intelligence Challenge (ConvAI2). In *The NeurIPS'18 Competition*. Springer, 187–208.

[10] Emily Dinan, Stephen Roller, et al. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *ICLR*.

[11] Eric N. Forsythand and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In *ICSC 2007*. 19–26.

[12] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[13] Molly E Ireland and James W Pennebaker. 2010. Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry. *Journal of Personality and Social Psychology* 99, 3 (2010), 549.

[14] Rricha Jalota, Priyansh Trivedi, et al. 2019. An Approach for Ex-Post-Facto Analysis of Knowledge Graph-Driven Chatbots - The DBpedia Chatbot. In *CONVERSATIONS*. 19–33.

[15] Raymond Li, Samira Ebrahimi Kahou, et al. 2018. Towards deep conversational recommendations. In *NeurIPS*. 9725–9735.

[16] Yanran Li, Hui Su, et al. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*. 986–995.

[17] Yinhan Liu, Myle Ott, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[18] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: A Novel Multi-Domain Information Seeking Dialogues Dataset. *arXiv preprint arXiv:1912.04639* (2019).

[19] Chen Qu, Liu Yang, et al. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations. In *SIGIR*. 989–992.

[20] Filip Radlinski, Krisztian Balog, et al. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *SIGdial*. 353–360.

[21] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *CHIIR*. 117–126.

[22] Abigail See, Aneesh Pappu, et al. 2019. Do Massively Pretrained Language Models Make Better Storytellers?. In *CoNLL*. 843–861.

[23] Abigail See, Stephen Roller, et al. 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. In *NAACL*.

[24] Johanne R. Trippas, Damiano Spina, et al. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *CHIIR*. 32–41.

[25] Marilyn A. Walker and Steve Whittaker. 1990. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In *ACL*. 70–78.