# Exploring Discourse Corpora Using Process Mining Techniques

Samantha Kent[1], Hans-Christian Schmitz[1]

[1]Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Fraunhoferstr. 20, 53343 Wachtberg, Germany

**Abstract**

In our paper we will introduce and discuss Process Mining (PM) as a means for conducting conversational analytic, linguistic and rhetorical investigations into discourse processes. PM is a technique for automatically deriving and further analyzing process models from event data. It is mainly applied for the analysis of Business Processes. We will argue that conversations and other kinds of discourse can be treated as processes too, and that PM enables us to systematically investigate large quantities of discourse transcripts. It would be possible to examine many different linguistic research questions, such as examining the conditions of turn-taking in dialogue or the pragmasemantics of discourse particles like English "well" or German "halt", among others.[1]

**Keywords**

Discourse Processing, Process Mining, Discourse Process Mining, Corpus Analysis, Information Extraction, Unsupervised Learning

## 1. Introduction

Process Mining (PM) is a young interdisciplinary research field that sits between machine learning and data mining on the one side and process modelling on the other [2]. The main difference to classic data-oriented types of analysis is that process mining focuses on the process as a whole, rather than just a specific aspect. Compared to process modelling, process mining relies on using real life raw data to model what is actually happening. Knowledge is extracted from raw data stored in event logs to discover, monitor and improve real processes. The data recorded by information systems can then be used to provide better insight into existing processes and the quality of process models can be improved.

There are three main types of PM: process discovery, conformance checking, and process enhancement [2]. Process discovery refers to the initial process in which a model is extracted from an event log. In conformance checking, data extracted from an event log are combined with an existing, predefined process model and the discrepancies are recorded as a diagnostic tool. It shows the difference between a model derived without data, i.e. what is supposed to happen, and the event log data, i.e. what is actually happening. Model enhancement combines

[1]The work in this paper is a continuation of our extended abstract [1]. It has been supported through funding from Philip Morris Impact as part of the Fraud Information Fusion Intelligence Project.

*Humanities-Centred AI (CHAI), Workshop at the 44th German Conference on Artificial Intelligence, September 28, 2021, Berlin, Germany*

samantha.kent@fkie.fraunhofer.de (S. Kent); hans-christian.schmitz@fkie.fraunhofer.de (H. Schmitz)

the previous types and extends or improves existing process models. It also allows the analysis of further aspects, such as time behaviour and consumed resources.

PM, which is sometimes also referred to as Business Intelligence, is applied in a wide variety of different fields. Examples of typical use cases are optimizing the processes in a hospital, order handling, or examining mortgage application processing in a bank. There are a number of prerequisites that are needed to analyse data using PM tools. In general, the data is stored in an event log, and contains activities that are further labelled with information, such as an activity label, an event time stamp, amount of resources needed, and other additional information. In order to conduct an analysis, at least the process ID, activity label and time stamp are needed.

In this paper, we used Process Mining techniques to explore the structure of discourse corpora. Section two introduces the concept of Discourse Process Mining. In section three, previous work is listed. Section four provides an overview of potential corpora that can be used for Discourse Process Mining, and the following section provides an exploratory example analysis of three of the corpora. Finally, we will illustrate the potential of Discourse Process Mining as a research method.

## 2. Discourse Process Mining

In this paper, we introduce Discourse Process Mining (DPM) and the idea that discourse structures can be modelled using PM techniques and tools. We argue that discourse, or more generally spoken and written conversation, contains a specific structure, a notion that is already explored in linguistics in e.g. Discourse Analysis (DA) and Speech Act Theory [3] [4]. Generally, dialogue proceeds in a linear manner, where the interlocutors exchange units of conversation that are functionally related to one another [5]. Successful discourse contains some type of organizational structure, this can be seen in turn-taking, opening and closing sequences of conversations, general conversational routines and repairs, and more specifically in adjacency pairs.

In DPM, we apply the techniques and tools used in traditional Process Mining to automatically extract structure from linguistic discourse corpora. We assume that DPM has a great potential for conducting investigations in both dialogue analysis and more general linguistics. Firstly, it would be extremely useful in optimizing processes where dialogue plays a central role, for example in customer service calls. It is imaginable that customer service centers keep transcripts of some of the conversations that also include time stamps and some type of customer satisfaction rating. It would therefore be possible to annotate the data and provide an overview of the general structure of satisfactory and not so satisfactory customer service calls. This is true for any conversational processes that are often formulaic, short and fulfil a specific purpose.

It would also be possible to examine many different linguistic research questions. These can range from more detailed linguistic questions pertaining to words in a discourse, or more general questions about conversational structure. For example, the use of specific discourse markers in a dialogue, such as the English “well” or “like” and the German “halt”, and the structural elements that precede or follow these markers. Another research question could relate to the use of speech acts between native speakers and foreign language learners. This question has already been extensively researched for the English language in particular, but a

structural approach using Discourse Process Mining could shed new light on an old research question. Furthermore, rhetorical analyses can reveal under which circumstances an argument is compelling and, thus, successful without necessarily being sound and valid. To us, the notion that dialogues can be interpreted as processes is quite compelling, and it might also be worth investigating whether monological texts can effectively be handled as processes too. Especially in the domain of academic writing, research into how to write an introduction contains specific structural guideline, and with DPM it would be possible to automatically analyse this type of structural information in academic papers. If corpora contain similar types of annotations, such research questions could not only be answered for a specific corpus or domain, but also provide a basis for a more general examination. Further details of specific corpora that may be suitable are given below.

There are a number of key requirements that are needed to conduct Discourse Process Mining. Let us assume that we are provided with a corpus of conversations. Each conversation will be treated as a separate process, each conversational move refers to a specific event, and each event has an annotated tag. The order of the events is equivalent to the order of the conversational moves in the corpus. In order to extract this information a corpus needs to be annotated using specific structural markers, such as speech or dialogue acts, and the corpus needs to be transformed in such a way that the annotations can be extracted and processed automatically. There are three key requirements that are needed to use PM tools to automatically process the data: each speech/dialog act needs a case ID, an activity and a timestamp. A case ID is the key of a process, it shows which process the utterance belongs to. The activity is the type of event, in other words the tag that belongs to the utterance. And finally, the timestamp is needed to ensure that the sequence utterances in the conversation is kept intact. Some tools will automatically adopt the input order and do not necessarily require a specific time stamp, however the inclusion of time stamps would enable a different type of analysis. Data preparation is key to be able to successfully extract meaningful information from a corpus.

## 3. Related Work

There has been some previous research using PM to explore dialogue structures. Most recently, Vakulenko et al. applied process mining techniques to discover patters in conversational transcripts of information-seeking dialogues [6]. These patterns are then used to develop an own model of conversation. The authors state that their model better represents conversations in this domain that previous models, because it better reflects the flow observed in real information-seeking conversations. While Vakulenko et al. focus on the analysis of a specific conversational domain, a handful of other studies focus on a specific corpus. Wang et al. described the application of PM techniques to analyze a corpus of online discussion threads from the Apple support forum [7]. Similarly, Compagno et al. developed a fine-grained corpus-independent classification of speech acts. They apply their annotations to a corpus of digital conversations extracted from the website Reddit and use PM tools to explore the written conversations [8]. Finally, Richetti et al. combine speech act theory and PM to automatically extract structure from customer service conversations [9].

## 4. Corpora

As is often the case with computational language analysis, one of the more challenging aspects of DPM is the availability of suitable data. As discussed above, a corpus needs to be annotated using an appropriate annotation scheme so that it can be transferred into an event log and automatically processed using DPM. We have found a number of different annotated corpora that would potentially be suitable to explore using DPM, some of which are listed below.

We have distinguished between two different categories of corpora, classic linguistic corpora that have been used in traditional linguistic corpus studies, and newer corpora that been collected for the purpose of examining language use for speech recognition systems. The Switchboard Dialogue Act corpus (SwDA) contains a collection of 1,155 five-minute telephone conversations between two participants [10]. It is annotated using the SWBD-DAMSL tag set. Once the initial data formatting is complete, DPM tools make it possible to automatically explore the conversational flow and the connections between the specific tags.

The ICSI meeting recorder dialogue act (MRDA) corpus contains about 180,000 hand-annotated dialogue act tags and accompanying adjacency pair annotations [11]. Interestingly, this corpus contains transcripts of meeting recordings, and provides a sample unconstrained speech in both a formal setting and of more casual conversation between multiple speakers. It has been annotated using the same tagset (SWBD-DAMSL) as in SwDA corpus, and is therefore often used as a comparison corpus.

In contrast to the two corpora above, the Spaadia corpus is a task-specific corpus of human-human train booking conversations. There are two different versions, and specifically the latest version annotated with the DART taxonomy seems to be suitable for DPM [12].

A more recent discourse corpus that has been developed for task-oriented dialogue modelling is the MultiWOZ corpus [13] [14]. It contains human-human written annotations that have been annotated with dialogue acts. It is the largest corpus containing just over 10.000 dialogues and spans different domains, including restaurant, hotel, police, and hospital, among others. Similarly, the Microsoft Dialogue Challenge consists of three domain specific corpora, taxi, restaurant and movie booking, collected for spoken dialogue modeling purposes [15]. The major difference is that the conversations in the corpus do not take place between two humans, but they rather simulate a conversation between a human and a conversational agent.

So far, all of the corpora introduced above have been monolingual English corpora. A corpus that is available in multiple different languages is the HCRC Map Task corpus [16]. This would allow for a comparative analysis of discourse structure in different languages.

The current discussion has centered around the processing of a selection of readily available discourse corpora. It would also be possible to create a corpus and annotate it so that it is possible to analyze the content using DPM. Given that hand-annotation is a very time consuming task, there is currently ongoing research into providing tools to automatically annotate corpora. On the one hand, this would drastically reduce the time it would to take to annotate a corpus and provide many new opportunities to explore previously unavailable discourse material. On the other hand, automatic annotation bears the risk that a further analysis rather leads to insights into the annotation algorithm and not into the original research question.

Whilst all of the corpora discussed above contain some type of dialog act annotation, they all differ in complexity, and are therefore more or less challenging to process using DPM. Based on

| | |
|---|---|
| ^h | A.1 utt1: {F Uh, } let's see. / |
| % | A.1 utt2: How [ about, + {F uh, } let's see, about ] ten years ago, / |
| qo | A.1 utt3: {F uh, } what do you think was different ten years ago from now? / |
| sv | B.2 utt1: {D Well, } I would say as, far as social changes go, {F uh, } I think families were more together. / |
| sv | B.2 utt2: [ They, + they ] did more things together. / |

**Figure 1:** An extract from the Switchboard Dialogue Act corpus. A and B are the speakers, utt stands for utterance, and the tags stand for hedge, interruption, open question and statement opinion respectively.

the experience we have with experimenting with the different corpora, we are exploring some of the potential research questions initially proposed in the following section.

## 5. Answering Potential Research Questions

The main goal in this paper is to introduce the concept of DPM and explore some of its potential uses. The research questions in this section serve as initial example analyses that illustrate the concept in general, rather than provide a detailed analysis. The analyses range from a structural dialogue level analysis, to a more detailed analysis of a specific linguistic phenomenon. We assume that there is a difference between task-specific corpora vs. unconstrained conversational corpora, and illustrate this by exploring different types of research questions based on the type of corpus. To start with, we applied DPM to the Switchboard corpus, which consists of unconstrained telephone conversations, albeit in a specific domain.

One of the challenges in this corpus is the large amount of variation. In total, there are 220 different tags in 42 different classes, and the conversations are quite long. This means there are many different possible combinations of tags within one conversation. Furthermore, the corpus was transcribed and annotated with a linguistic analysis mind. It therefore contains a detailed analysis of hesitations, overlaps, and other features that are not necessary relevant for a more pragmatic structural analysis. An example of the transcript can be found in figure 1. When processing these dialogues using DPM, it quickly became apparent that editing and carefully selecting the data was crucial to gaining meaningful insights for a corpus with so much variance. Figure 2 shows the process map of a partial structural corpus analysis, and the outcome can be described as a spaghetti model [2]. While it is easy to input the data and automatically create these models, they are difficult to interpret and not suitable for this type of research. We therefore conclude that it would be more suitable as a basis for exploring more specific linguistic research questions such as the pragmatic use of the discourse marker "well". Whilst this analysis is beyond the scope of this paper, we plan to address these questions in future work.

In contrast, the results from the automatic structural analysis of a task specific corpus, such as the train booking dialogue from the Spaadia corpus show great potential. Figure 3 shows the process map of an analysis of 35 train human-human train booking conversations. Because there

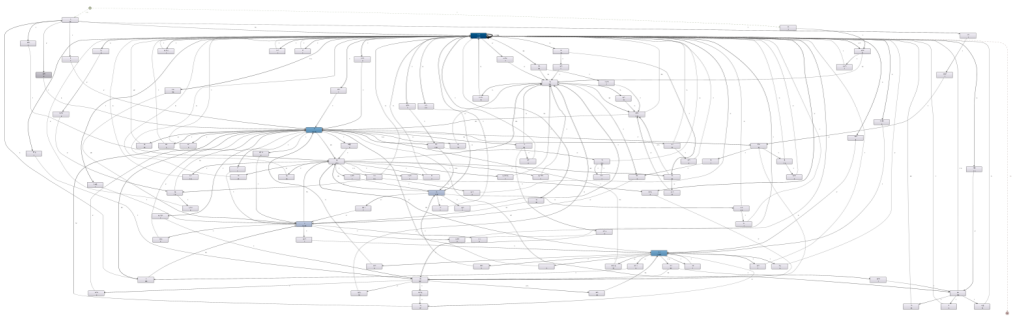

**Figure 2:** The process map, a so-called spaghetti model, from the SwDA corpus. The data was analyzed using the Process Mining tool Disco by Fluxicon.

is much less variance in a task-specific dialogue, the process map is much more condensed and it shows some of the general structure of the conversations. Starting from the top of the process, almost all of the conversations start with a greeting and the speakers identify themselves. In the middle of the conversation there is a split into different paths. Interestingly, the model also shows that there are a number of conversations that contain multiple bookings or requests, the dark arrow shows that the process starts again with an information request rather than ending directly.

To further illustrate the concept of DPM, we also analyzed the Microsoft Dialogue Challenge corpus, specifically because it is also a constrained task-orientated corpus, but unlike the Spaadia corpus, it involves a human and an conversational agent. In total, the corpus contains the transcriptions of about 3000 conversations where a system user interacts with a movie booking agent. Figure 4 shows that DPM enables the quick analysis of the overall structure of the dialogues in the corpus. The process map shows some basic information, for example that the two most used dialogue acts are request and inform, and that they often follow one another. More interestingly, it also shows that the conversations often do not contain a more traditional ending to a conversation, as only 1169 occurrences of thanking, and consequently only 50 instances of welcome, were found in the data.

It is important to note that generalizations about conversational structure only hold true for the specific corpus that is being researched. Especially in the case of the examples above, the dialogues between a human-agent seem to be more structured than between human-human interlocutors, as conversational turns are dictated by the system used by the agent. Taxi bookings, or bookings more generally, seem to show more variance, as in the Spaadia corpus. However, Discourse Process Mining can be used as a tool to further explore conversational structure in a way that would have been difficult to achieve so rapidly using more traditional methods. It provides an automated process to gain quick insights into different types of conversational corpora.

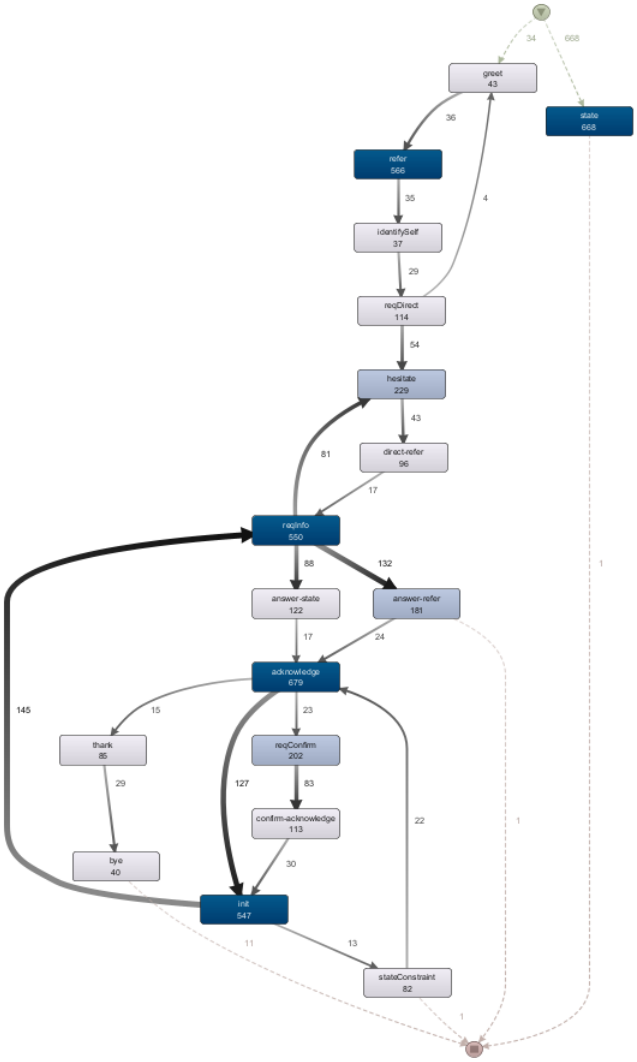


**Figure 3:** The process map based on the Spaadia corpus. The data was analyzed using the Process Mining tool Disco by Fluxicon.

## 6. Conclusion

The aim of this paper was to introduce the idea of applying automatic Process Mining (PM) techniques to the analysis of discourse phenomena. A discourse is a sequence of speech acts or conversational moves. From a corpus of conversations, in other words different discourse sequences, a comprehensive discourse model can be derived. Intuitively, there is a difference between conducting Discourse Process Mining (DPM) on corpora containing unrestrained dialogue versus highly constrained, task-oriented dialogue. Our initial research supports this hypothesis as was demonstrated using the Switchboard Dialogue Act Corpus, the Spaadia and the Microsoft Dialogue Challenge corpus.

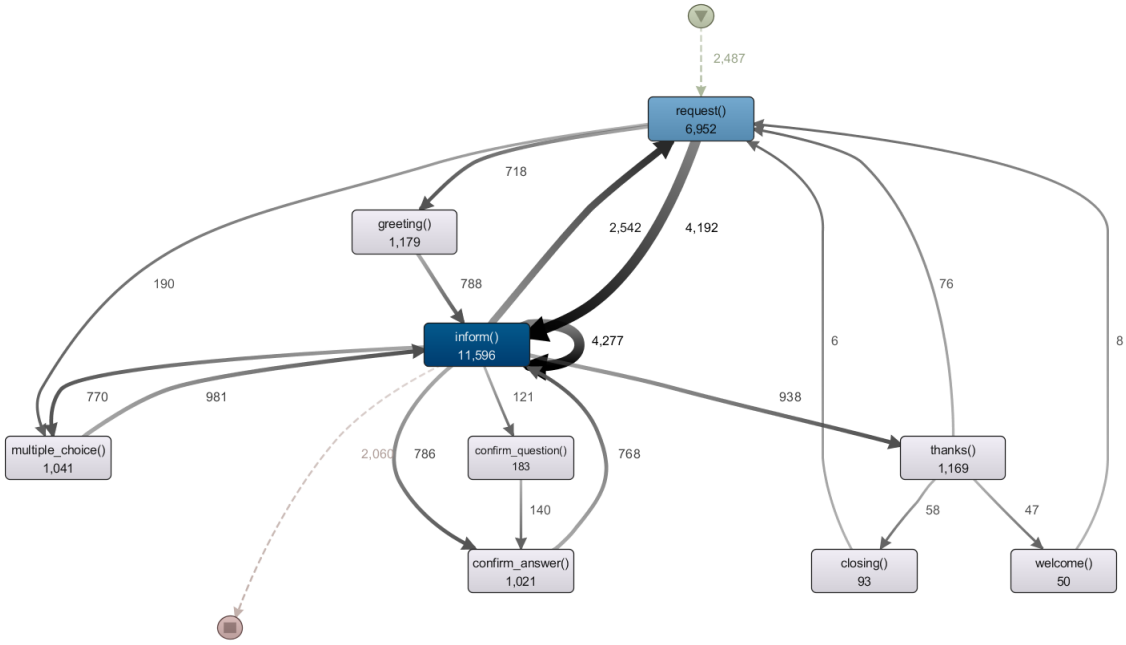


**Figure 4:** The analysis of the movie-booking task in the Microsoft Dialogue Challenge corpus. The process map was created using the Process Mining tool Disco by Fluxicon.

There are various tools for conducting Process Mining and, therefore, Discourse Process Mining. For our initial analyses, we used the tool Disco by Fluxicon (https://www.fluxicon.com/), which has a high usability. There is also a very active PM community. Therefore, experiments in DPM are easy to conduct. One major constraint is the availability of a suitable discourse corpus that (a) contains a sufficient number of phenomena to be investigated and (b) is reliably annotated. As is often the case in data-driven linguistics, the availability of data can be a problem, and tools for automatic annotation of suitable corpora might be of use in this case.

In general, we assume that discourse process models can effectively support the further investigation of discourse structure, as well as the speakers' means to control discourse, in particular during conversations with multiple participants. In future work we would like to explore some of the potential research questions introduced in this paper, in particular the examination of discourse particles, such as "well" or the German "halt" in large unconstrained speech corpora.

## References

[1] S. Kent, H.-C. Schmitz, Discourse process mining, Humanities-Centred AI (CHAI) (2021). URL: https://doi.org/10.25592/uhhfdm.9672.

[2] W. M. P. van der Aalst, Process Mining: Data Science in Action, 2 ed., Springer, Heidelberg, 2016. doi:10.1007/978-3-662-49851-4.

[3] J. L. Austin, How to Do Things with Words, Harvard University Press, Cambridge, MA, 1962.
[4] J. R. Searle, Speech Acts, Cambridge University Press, Cambridge, UK., 1969.
[5] E. A. Schegloff, H. Sacks, Opening up closings, Semiotica 8 (1973) 289–327.
[6] S. Vakulenko, K. Revoredo, C. D. Ciccio, M. de Rijke, QRFA: A data-driven model of information-seeking dialogues, CoRR abs/1812.10720 (2018). URL: http://arxiv.org/abs/1812.10720. arXiv:1812.10720.
[7] G. A. Wang, H. J. Wang, J. Li, A. S. Abrahams, W. Fan, An analytical framework for understanding knowledge-sharing processes in online qa communities., ACM Trans. Manag. Inf. Syst. 5 (2015) 18:1–18:31. URL: http://dblp.uni-trier.de/db/journals/tmis/tmis5.html#WangWLAF15.
[8] D. Compagno, E. Epure, R. Deneckère, C. Salinesi, Exploring digital conversation corpora with process mining, Corpus Pragmatics 2 (2018). doi:10.1007/s41701-018-0030-6.
[9] P. H. P. Richetti, J. C. de A. R. Gonçalves, F. A. Baião, F. M. Santoro, Analysis of knowledge-intensive processes focused on the communication perspective., in: J. Carmona, G. Engels, A. Kumar (Eds.), BPM, volume 10445 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 269–285. URL: http://dblp.uni-trier.de/db/conf/bpm/bpm2017.html.
[10] D. Jurafsky, E. Shriberg, D. Biasca, Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, Technical Report Draft 13, University of Colorado, Institute of Cognitive Science, 1997.
[11] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, H. Carvey, The icsi meeting recorder dialog act (mrda) corpus., in: M. Strube, C. L. Sidner (Eds.), SIGDIAL Workshop, The Association for Computer Linguistics, 2004, pp. 97–100. URL: http://dblp.uni-trier.de/db/conf/sigdial/sigdial2004.html.
[12] M. Weisser, Dart – the dialogue annotation and research tool, Corpus Linguistics and Linguistic Theory 12 (2016). doi:10.1515/cllt-2014-0051.
[13] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, U. Stefan, R. Osman, M. Gašić, Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
[14] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, J. Chen, Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines, in: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020, 2020, pp. 109–117.
[15] X. Li, S. Panda, J. Liu, J. Gao, Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems, CoRR abs/1807.11125 (2018). URL: http://arxiv.org/abs/1807.11125. arXiv:1807.11125.
[16] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, R. Weinert, The hcrc map task corpus, Language and Speech 34 (1991) 351–366. URL: https://doi.org/10.1177/002383099103400404.