

A Comparative Analysis of Crowdsourced Natural Language Corpora for Spoken Dialog Systems

Patricia Braunger¹, Hansjörg Hofmann¹, Steffen Werner¹, Maria Schmidt²

¹Daimler AG, Sindelfingen, Germany

²Karlsruhe Institute of Technology, Karlsruhe, Germany

{patricia.braunger, hansjoerg.hofmann, steffen.s.werner}@daimler.com, maria.schmidt@kit.edu

Abstract

Recent spoken dialog systems have been able to recognize freely spoken user input in restricted domains thanks to statistical methods in the automatic speech recognition. These methods require a high number of natural language utterances to train the speech recognition engine and to assess the quality of the system. Since human speech offers many variants associated with a single intent, a high number of user utterances have to be elicited. Developers are therefore turning to crowdsourcing to collect this data. This paper compares three different methods to elicit multiple utterances for given semantics via crowd sourcing, namely with pictures, with text and with semantic entities. Specifically, we compare the methods with regard to the number of valid data and linguistic variance, whereby a quantitative and qualitative approach is proposed. In our study, the method with text led to a high variance in the utterances and a relatively low rate of invalid data.

Keywords: natural language data, crowdsourcing, elicitation methods

1. Introduction

Early spoken dialog systems (SDS) supported only a finite set of speech commands. Thanks to statistical methods in the automatic speech recognition (ASR) recent systems such as Apple's Siri¹ have been able to recognize freely spoken user input in restricted domains. These methods require a high number of spoken utterances to train the speech recognition engine. In order to train the speech recognition engine and to assess the quality of a SDS which is able to interpret any spoken user utterance associated with a given intent, it is necessary to collect data that reflects real usage. The elicitation of real user data is a challenge within the development process of such systems. From a commercial point of view, data collection for SDS should meet the following requirements:

- Allow for fast elicitation
- Apply to different intents and domains
- Reflect representative user input
- Include a high rate of usable data
- Cover high linguistic variance.

Since development cycles of SDS are normally short and much data is needed, developers are turning to crowdsourcing (Eskenazi et al., 2013). Our goal is to investigate the use of crowdsourcing methods with regard to the requirements above. In order to collect real system interaction data, one should choose elicitation methods which do not bias the participants. However, a high number of valid utterances should be generated.

In this paper, we collected natural language speech utterances via crowdsourcing with the goal of comparing three different elicitation methods. The participants were asked

to perform tasks such as entering the address into the navigation application. The elicitation methods differed in the way how the tasks were presented to the participants, by means of:

- Pictures
- Semantic entities
- Text.

Specifically, we compare the methods with regard to the number of valid data and linguistic variance, whereby a quantitative and qualitative analysis approach is proposed. The focus of the data we collected was on the usage as test data in the in-car domain. However, these elicitation methods are applicable to other domains and they could also be used to acquire training data, e.g. for the training of statistical language models. In this paper, we will also propose a guideline for future crowdsourced data collections for use in SDS.

The paper at hand is structured as follows. The next section gives an overview on previous studies on this research topic. Section 3 describes the data collection setup. Section 4 presents the results and discusses the findings. Section 5 outlines several lessons learned and in Section 6, conclusions are drawn.

2. Related Work

Crowdsourcing, a relatively common approach, consists of outsourcing tasks to a broad external group of nonexperts via Internet platforms (Eskenazi et al., 2013). The benefits of crowdsourcing can include rapidness, low costs, the individuality and creativity of people, the diversity of opinions and the required quantity, depending on the field of application (Hosseini, 2015; Vukovic, 2009). Beside the benefits, crowdsourcing is often criticized as producing poor quality because it is difficult to control the work quality and the status of the workers (Eskenazi et al., 2013).

¹ <http://www.apple.com/ios/siri/>

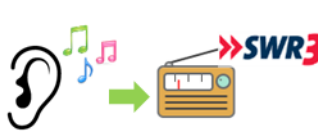
Pictures	Semantics	Text
	<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">hören "listen"</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Sender "radio station"</div> <div style="border: 1px solid black; padding: 5px;">SWR3</div>	<p>Stellen Sie sich vor, Sie fahren eine lange Strecke mit dem Auto. Für ein bisschen Abwechslung möchten Sie den Radiosender SWR3 einstellen. Was sagen Sie?</p> <p>"Imagine that you are travelling by car. For a little change you would like to switch to the radio station SWR3. What would you say?"</p>

Figure 1: Different task presentations for the elicitation of speech utterances.

However, in the past years, the speech processing communities have realized that crowdsourcing is a possible solution to their strong need for speech data (Eskenazi et al., 2013). Crowdsourcing has been used for speech acquisition (McGraw et al., 2010), speech transcription and annotation (Jyothi et al., 2015; Sabou, 2014), evaluation of speech technology (Yang et al., 2010) and paraphrase generation (Burrows et al., 2013). Manuvinakurike et al. (2015) show the potential of crowdsourced spoken dialog data to lower costs and facilitate the design and evaluation of SDS.

Eliciting linguistic variants that correspond to a given semantics seems to be a relatively young field of research in the context of crowdsourcing. Work in this field of research has been done by Yang Wang et al. (2012), where the authors present three methods to elicit natural language exclusively for semantic frames with slots and values, e.g. *FindRestaurant(City=Seattle; Cuisine=Chinese)*. For semantic interpretations such as given in the example the authors create sentences, scenarios and lists which describe the intent to the crowd workers. In the sentence-based method a corresponding natural language sentence is given. In the scenario-based method a scenario is given via multiple sentences and the list-based method presents the slots and values in the form of a list. Their evaluation focuses on the distribution over possible slot orderings. Since the elicitation and analyzing methods they used, show some limitations to elicit data for in-car SDS, we argue for methods that satisfy the requirements that are mentioned above. Specifically, our methods are applicable to different kinds of commands: Commands with one slot like *"Call John Smith"*, commands with more than one slot like *"Navigate to 11 Main Street in Springfield"* and also simple commands like *"Next station"*.

3. Data Collection Setup

With the help of the German crowdsourcing company Clickworker², we asked the crowd worker community to give voice input to a fictitious spoken dialog system. In a first step, the crowd workers had to verbally input and record one utterance for seven tasks such as entering an address in the navigation application. In the second step, they had to transcribe the utterance themselves. The elicitation methods we have chosen to present the task can

be applied to different intents in different domains. We investigated the following task presentation methods: pictures, semantics and text (see Figure 1).

In the elicitation method with pictures task presentation, the participants were shown a picture depicting the task they should perform. In the semantics task presentation, the participants get presented three semantic entities. As for the text presentation, the participants were presented a few lines describing the situation they are in and the actions they should perform. Within each of the presentation methods we asked for user speech input for seven tasks typically performed in the car:

- 1) Listen to radio station SWR3
- 2) Play Michael Jackson Greatest Hits
- 3) Next Shell gas station
- 4) Navigate to Stieglitzweg 23 in Berlin
- 5) Call Barack Obama on his mobile phone
- 6) Set temperature to 23°C
- 7) Send a text message to brother

That means, the presentations in task 1 explain that the user of the SDS wants to listen to a certain radio station.

4. Evaluation

The most important property of natural language compared to simple commands is a variable wording and a flexible sentence construction, i.e. flexible constituent order or different sentence types. In order to evaluate the proposed methods, we investigate the differences between the most frequently used utterances and particularly, the differences on word and on sentence level. Different studies, e.g. Bernsen et al. (1998), report from priming effects when using text-based scenario descriptions. The pictures method does not bias the subjects by putting words into their mouths. As the pictures method favors the use of different words and sentence constructions, we take the utterances from the pictures method as a reference to detect priming effects. Before investigating the linguistic variance, we first summarize the characteristics of the corpora and compare the rate of valid data.

4.1 The Corpora

For each of the seven tasks we collected 1,080 speech utterances for every presentation method via crowdsourcing. The 3,240 crowd workers were German native speakers. 90% of the crowd workers were between 18 and 35 years old, 8% of the crowd workers was up to

² <http://www.clickworker.com/>

55 years old and 2% was aged over 55 years. 60% of the crowd workers were male and 40% female.

Preliminary investigations of the collected corpora presented in Schmidt et al. (2015) focused on techniques for recognizing errors and eliminating faulty data sets. In addition, the characteristics of the resulting data sets were described in Schmidt et al. (2015). In this paper, the data collection is evaluated in detail regarding the validity of the utterances and the linguistic variance. In order to analyze the corpora in detail, the few pre-processing steps from Schmidt et al. (2015), such as spell-checking and text normalization (e.g. lowercasing and eliminating punctuation) were applied to the collected utterances.

4.2 Valid Data

Based on the normalized data we examined the rate of valid utterances. The data was semi-automatically filtered by certain semantic keywords. For every task obligatory content words were determined that have to be named in any way, cf. Schmidt et al. (2015). If at least one of the determined words occurs, the utterance is classified as valid. If none of the semantic keywords occurs, the utterance is classified as not valid. However, there still remain a few faulty utterances among the valid ones because the keywords may occur in utterances where the participants misunderstood the task. In section 5 we discuss possible solutions to avoid faulty data. Table 1 displays the percentage of the valid utterances in which the obligatory content words were named.

	Pictures	Semantics	Text
Valid utterances	72%	92%	87%

Table 1: Valid utterances, cf. Schmidt et al. (2015).

In the pictures method we detect significantly fewer valid utterances than in the semantics method ($p < 0,01$) and in the text method ($p < 0,02$). There is no significant difference in the semantics and the text method ($p > 0,05$). As for the pictures presentation, for instance, we detect quite a big difference in the rate of valid data between the tasks; see Coefficient of Variation (CoV) in Figure 2.

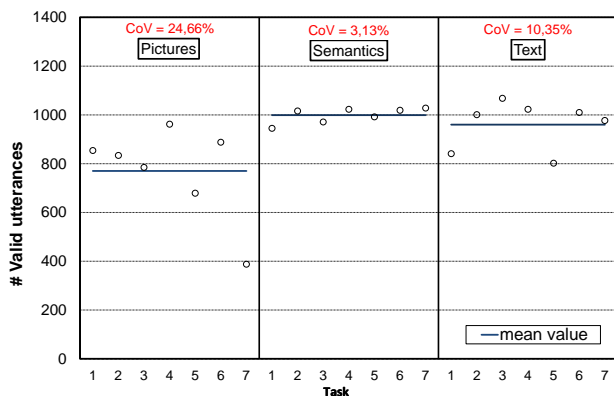


Figure 2: Differences in the rate of valid utterances.

In task 5 of the pictures presentation method for example, five of the ten most frequently used utterances are classified as not valid. Each of the ten most frequently used utterances in the semantics presentation method is classified as valid.

Our further comparative analysis operates on the valid data set, in order to find out the optimal presentation method to get a high linguistic variance in the collected speech utterances.

4.3 Most Frequently Used Utterances

Concerning the most frequently used utterances, we provide answers to the following questions:

- Are there preferred utterances?
- Do the preferred utterances reflect priming effects?

Some preferences seem to exist, e.g. in task 1. Both, the utterances from the pictures and the text method show a distance between the most and the second and third frequent utterance, see Table 2.

Pictures	Utterance	Text
11,2%	Radio SWR3 "Radio SWR3"	8,4%
4,6%	Radiosender SWR3 "Radio station SWR3"	3,4%
	Sender SWR3 einstellen "Switch to radio station SWR3"	
4,1%	Ich möchte SWR3 hören "I would like to listen to SWR3"	3,3%

Table 2: Most frequent utterances, task 1.

In addition, the most frequent utterance from the pictures method is identical to the most frequent utterance from the text method over four tasks. Some of the differences between the pictures and the text method are due to priming effects, e.g. the discrepancy in the second frequent utterance, where the verb "einstellen" (eng. "tune") was given in the text description. The most frequently used utterances in the semantics presentation method differ compared to the pictures method, see Table 3. They seem to reflect strong priming effects. The semantics presentation method gives one of the semantic entities in form of a verb over all tasks, e.g. "listen" in task 1. Since nearly each of the ten most frequent utterances per task consists of the given verb, one can conclude that this is an effect of priming.

Utterance	Occurrence
Ich möchte den Sender SWR3 hören "I would like to listen to the station SWR3"	20,0%
Sender SWR3 hören "Listen to the station SWR3"	16,1%
Sender hören SWR3 "Listen to the station SWR3"	4,4%

Table 3: Most frequent utterances in the semantics method, task 1.

Pictures		Semantics		Text	
Word	Occurrence	Word	Occurrence	Word	Occurrence
swr3	838	swr3	946	swr3	838
radiosender	508	sender	737	sender	408
radio	436	hören	685	radio	275
sender	194	radiosender	270	radiosender	241
hören	148	radio	134	einstellen	135
suchen	103	schalten	44	suchen	97
schalten	97	suchen	36	hören	88
einstellen	72	einstellen	31	schalten	70
spielen	63	spielen	30	spielen	44
einschalten	46	einschalten	20	wechseln	25

Table 4: Number of different content words, task 1.

4.4 Word Level

As for the word level analysis, we identified the most frequent words of each task on the basis of the valid data set. We have concentrated on different lexical content words regardless of their morphological surface form. Exemplarily, Table 4 displays the distribution of the used words in task 1 for the pictures, semantics and text method. The other tasks show similar distributions.

One can see that the participants used similar words. The corpora differ in the distribution of the words used. The three most frequent words in the semantics method are identical to the given semantic entities over all tasks, for an example see Table 4. “SWR3”, “Sender” (eng. “station”) and “hören” (eng. “listen”) are given in the task presentation. Furthermore, the significant increase in the frequency is an indicator of a lower lexical variance. As for the text method, we detect some priming effects, too. “Einstellen” (eng. “tune”) which is the most frequent verb in the text method was given in the text description but less frequently used in the pictures method.

4.5 Sentence Level

The sentence level analysis included Part-of-Speech (POS) Tagging with the Tree Tagger (Schmid, 1994). We semi-automatically clustered the POS sequences. The clustering gives an impression of the sentence constructions people use speaking freely to the system. Table 5 displays the seven resulting sentence constructions as outcome from the clustering. We found that there is quite a big difference in the manner people speak to the system. They use a command style, e.g. “Radio SWR3”, as well as full natural sentences, see Table 5. We conclude that if people speak freely to a SDS, they mostly use the imperative style. Next, we compared the elicitation methods with regard to the preferred sentence constructions. The different sentence constructions appear within all three elicitation methods. Figure 3 displays the distribution of the sentence constructions over all tasks.

The most common sentence constructions are the *imperative*, *command style* and *command style & infinitive*. The *imperative* and the *command style & infinitive* were by far the most frequently used sentence constructions over all tasks and methods. However, it is striking that the *command style & infinitive* is the most preferred sentence construction in the semantics method, whereas the *imperative* is the most preferred sentence construction in the pictures and in the text method.

Sentence construction	Example
Interrogative & pronoun	Wo ist die nächste Shell-Tankstelle? “Where is the nearest Shell gas station?”
Interrogative & modal verb	Kannst du den Sender SWR3 einstellen? “Could you switch to the radio station SWR3?”
Indicative	Ich suche den Radiosender SWR3 “I’m searching for the radio station SWR3”
Indicative & modal verb	Ich möchte/will SWR3 hören “I would like to listen to SWR3”
Command style & infinitive	SWR3 einstellen “No corresponding syntax existing in English”
Command style	Radio SWR3 “Radio SWR3”
Imperative	Spiele SWR3 “Play SWR3”

Table 5: Sentence constructions that appear in the corpora.

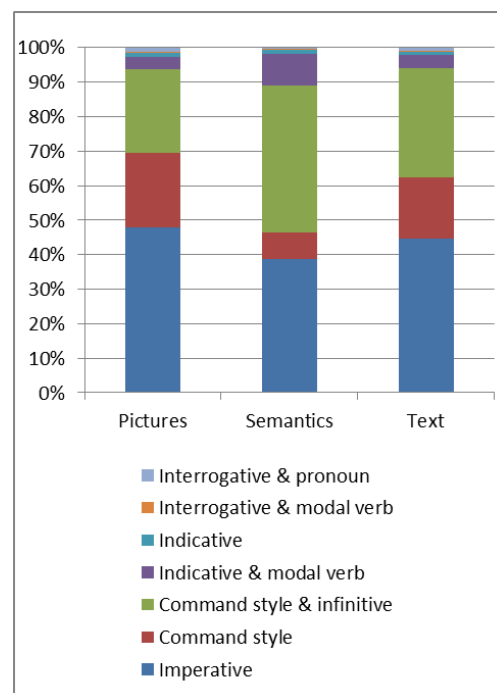


Figure 3: Distribution of sentence constructions.

Again, this seems to be an effect of priming. Since the semantic entities present an infinite verb, the participants tend to produce sentences that contain an infinite verb form. One can see another priming effect in Figure 4. Figure 4 displays the four most frequent sentence constructions in task 1.

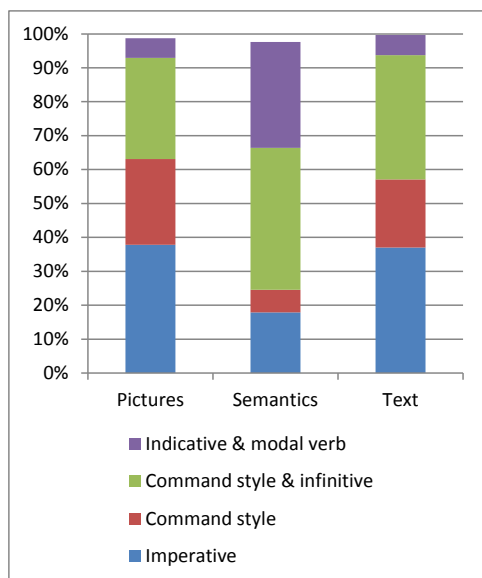


Figure 4: Most frequent sentence constructions, task 1.

The second frequent sentence construction in the semantics method is the *indicative & modal verb* which is not frequently used in the pictures and in the text method. The effect is due to the verb “hören” (eng. “listen”), which is given in the semantics presentation method. It is not possible to use the verb “hören” in the imperative mood in this context. A voice command like “Höre SWR3!” (eng. “Listen to SWR3!”) makes no sense. The participants seem to use the *indicative & modal verb* sentence construction instead of the *imperative*.

Depending on the nature of the task certain sentence constructions are preferred as Figure 4 and 5 show.

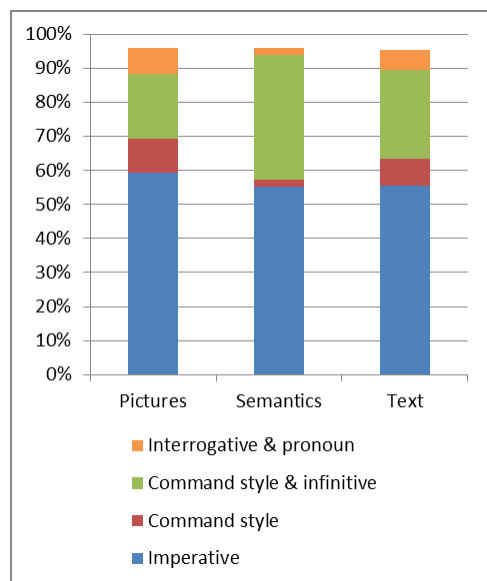


Figure 5: Most frequent sentence constructions, task 3.

Whereas the *indicative & modal verb* is among the four most frequent sentence constructions in task 1, the participants prefer the *interrogative & pronoun* construction in task 3. The *imperative* is used more frequently in task 3 and the *command style* is used more frequently in task 1. Figure 5 shows again similarities between the text method and the pictures method.

5. Limitations and Recommendations

Based on our findings, for future data collections we recommend making use of the text presentation method. We advise against making use of the semantics method. The method elicited a high number of valid data but the data show strong priming effects. Whereas the pictures method doesn’t bias the participants but produces a high number of faulty data, the text presentation method is a good compromise between a high rate of valid utterances and a great linguistic variance. The text presentation method facilitates task descriptions for different intents and domains and minimizes misunderstandings. However, one should be aware of potential priming effects. In order to avoid priming, we recommend paraphrasing technical terms. Parameters like address parts should be given in multiple short sentences, e.g. “Imagine that you would like to visit your friend who lives in Springfield. The street is the Main St. The house number is 11.” Thus, the participants are less biased in the parameter ordering and the use of linking elements.

In order to increase the number of valid utterances within the text method, we recommend the following procedure. We identified different causes for faulty data. We found technical causes, task misunderstandings and wrong transcriptions. We recommend filtering users with a bad audio setup. As an example, Manuvinakurike et al. (2015) made the users to listen to an audio file and transcribe it. A wrong transcription disqualified the user. The participants then had to speak three determined sentences in their microphone. An ASR transcribed the spoken words. If the participant had no word right in each sentence, the participants were disqualified. Task misunderstandings can be prevented by clear task descriptions and examples in the instructions. The self-transcription of the participants’ own speech shows some limitations as well. In some cases, the transcription doesn’t match the spoken words. For future data collections we therefore suggest that participants do not transcribe their own audio recordings.

6. Conclusion

This paper presented a comparative analysis of three different data elicitation methods via crowdsourcing. In order to find out the best elicitation method for natural language data collections, we analyzed the most frequent utterances, the differences on sentence level and on word level. We showed great similarities between the text presentation and the pictures presentation method. We also showed that the significant differences between the semantics presentation and the pictures presentation are an effect of priming. On the basis of the analysis findings we suggested making use of the text presentation method. Although some minor priming exists, it is a good compromise between a high rate of valid data, the linguistic variance and the possibility of creating very specific tasks

for different types of commands. In order to make a statement on the representativeness, some further research is needed. One should compare younger groups' utterances to elder groups' utterances. This study was targeted at German native speakers. In order to evaluate the feasibility of the proposed approach for other languages, we aim to assign the data collection approach to other languages, European and non-European.

7. References

- Bernsen, N. O., Dybkjaer, H. and Dybkjaer, L. (1998). Designing Interactive Speech Systems: From First Ideas to User Testing. Berlin: Springer.
- Burrows, S., Potthast, M. and Stein, B. (2013). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *ACM Transactions on Intelligent Systems and Technology, Vol. 4, Issue 3*, pp. 1--22.
- Eskenazi, M., Levow, G., Meng, H., Parent, G. and Suendermann, D. (2013). Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment. John Wiley & Sons.
- Hosseini, M., Shahri, A., Phalp, K., Ali, R. (2015). Recommendations on Adapting Crowdsourcing to Problem Types. *Proceedings of the IEEE 9th International Conference on Research Challenges in Information Science*, pp. 423--433.
- Jyothi, P., Hasegawa-Johnson, M. (2015). Transcribing continuous speech using mismatched crowdsourcing. *Proceedings from Interspeech*, pp. 2774--2778.
- Manuvinakurike, R., Paetzel, M., DeVault, D. (2015). Reducing the Cost of Dialogue System Training and Evaluation with Online, Crowd-Sourced Dialogue Data Collection. *Proceedings of the Workshop on Semantics and Pragmatics of Dialogue 2015*.
- McGraw, I., Lee, C., Hetherington, L., Seneff, S., Glass, L. (2010). Collecting voices from the cloud. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 1576--1583.
- Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 859--866.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the international conference on new methods in language processing, volume 12*, pp. 44--49.
- Schmidt, M., Müller, M., Wagner, M., Stüker, S., Waibel, A., Hofmann, H. and Werner, S. (2015). Evaluation of Crowdsourced user input data for spoken dialog systems. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 427--431.
- Vukovic, M. (2009). Crowdsourcing for enterprises. *Proceedings of World Conference on Services-I*, pp. 686--692.
- Yang Wang, W., Bohus, D., Kamar, E. and Horvitz, E.. (2012). Crowdsourcing the acquisition of natural language corpora: methods and observations. *Proceedings of the IEEE Workshop on Spoken Language Technology*, pp. 73--78.
- Yang, Z., Li, B., Zhu, Y., King, I., Levow, G. and Meng, H. (2010). Collection of user judgments on spoken dialog system with crowdsourcing. *Proceedings of the IEEE Workshop on Speech and Language Technologies*.