2017

Crowdsourcing a Parallel Corpus for Conceptual Analysis of Natural Language

Jamie C. Macbeth
Fairfield University, jmacbeth@smith.edu

Sandra Grandic
Fairfield University

Proceedings of the Fifth Conference on
Human Computation and Crowdsourcing
(HCOMP 2017)

Crowdsourcing a Parallel Corpus for Conceptual Analysis of Natural Language

Jamie C. Macbeth, Sandra Grandic
Department of Electrical and Computer Systems Engineering
Fairfield University
1073 North Benson Road
Fairfield, Connecticut 06824

Abstract

Computer users today are demanding greater performance from systems that understand and respond intelligently to human language as input. In the past, researchers proposed and built conceptual analysis systems that attempted to understand language in depth by decomposing a text into structures representing complex combinations of primitive acts, events, and state changes in the world the way people conceive them. However, these systems have traditionally been time-consuming and costly to build and maintain by hand.

This paper presents two studies of crowdsourcing a parallel corpus to build conceptual analysis systems through machine learning. In the first study, we found that crowdworkers can view simple English sentences built around specific action words, and build conceptual structures that represent decompositions of the meaning of that action word into simple and complex combinations of conceptual primitives. The conceptual structures created by crowdworkers largely agree with a set of gold standard conceptual structures built by experts, but are often missing parts of the gold standard conceptualization. In the second study, we developed and tested a novel method for improving the corpus through a subsequent round of crowdsourcing; In this "refinement" step, we presented only conceptual structures to a second set of crowdworkers, and found that when crowdworkers could identify the action word in the original sentence based only on the conceptual structure, the conceptual structure was a stronger match to the gold standard structure for that sentence. We also calculated a statistically significant correlation between the number of crowdworkers who identified the original action word for a conceptual structure, and the degree of matching between the conceptual structure and a gold standard conceptual structure. This indicates that crowdsourcing may be used not only to generate the conceptual structures, but also to select only those of the highest quality for a parallel corpus linking them to natural language.

Introduction

Building systems to understand and respond to natural language input has been a goal of artificial intelligence research for decades, and recently researchers have applied novel machine learning techniques to vast text corpora to solve problems posed by natural language processing (NLP). But, as the authors of a recent survey of NLP research assert, "the truly difficult problems of semantics, context, and knowledge will probably require new discoveries," (Hirschberg and Manning 2015) and solutions to many problems in NLP still appear to be out of reach using existing corpora. Another rich tradition of work in cognitive artificial intelligence strives for human-like performances on language comprehension tasks by performing analyses of text that are driven by conceptual representations rather than syntax, phonology, and morphology. *Conceptual analysis* systems attempt to transform a text into a non-linguistic representation that reflects how humans conceive of physical and social situations represented by the language, to support the sort of memory retrieval and inference processes that humans perform when understanding language in depth.

A conceptual analyzer requires a mapping between lexical items—words and phrases—and the non-linguistic conceptual structures that form the elements of subsequent analyses, a subsystem traditionally built by hand. As with many efforts in symbolic artificial intelligence, the development of conceptual analysis systems must confront the knowledge engineering problem: the impracticality of building and maintaining such systems at scale manually, especially when machine learning technologies are available. However, to build conceptual analysis systems via machine learning will require at least an initial kernel of non-linguistic conceptual structures tied to language, which are unlikely to be gleaned from existing corpora (Schuler 2005; Miller 1995), because they comprise commonsense knowledge understood by all competent language users that typically goes unsaid.

For many years, crowdsourcing (Howe 2006) has been used to annotate datasets for natural language processing based on machine learning. More recently, crowdsourcing has even demonstrated promise for collecting narrative intelligence knowledge (Li et al. 2012). But there are great potential challenges to crowdsourcing a corpus for machine learning-based conceptual analysis: building conceptual structures is an abstract and complex process, even for experts, and crowdworkers likely will not have background knowledge of the conceptual representation—either its primitive elements, or the connective elements that allow one to build larger structures.

This paper presents studies of methods for leveraging crowdsourcing to develop a corpus for conceptual analysis.

We find that crowdworkers can build both simple and complex conceptual structures in a language-free representation called Conceptual Dependency. The conceptual structures they build are based on their understandings of simple sentences in English, and they largely matched and agreed with a set of gold standard conceptual structures created by experts. We also developed and tested a novel method for improving the corpus of conceptual structures through a subsequent round of crowdsourcing. We found that when we presented decomposed conceptual structures to a second set of crowdworkers, they could often "recompose" them and identify the action word in the original sentence, and this indicated that the conceptual structure was a strong match to the gold standard structure for that word. We calculated a statistically significant correlation between the fraction of times crowdworkers could recompose the conceptual structure and determine the original action word, and the degree of matching between the conceptual structure and a gold standard conceptual structure created by experts. This demonstrates that crowdsourcing can be used to generate a parallel corpus of conceptual structures tied to lexical items, and it can be used refine the corpus to achieve a quality comparable to that provided by experts knowledgeable in the representation.

Background and Prior Work

In the last century, experimental psychologists and psycholinguists who studied language behavior showed that when human listeners understand discourse, they quickly forget its grammatical form or syntax (Sachs 1967), and whether a particular meaning was conveyed by a noun or verb (Johnson-Laird, Robins, and Velicogna 1974). Other related research found that humans tend to construct mental models (Johnson-Laird 1983) and imagery (Paivio 1971) rather than ontologies or structures in classical logic (Rosch 1975; Wason and Johnson-Laird 1972) in their language understanding processes. Based on these insights, many artificial intelligence researchers of the same period were motivated by psychological and cognitive validity instead of grammar- and syntax-directed analyses when building systems to understand the meaning behind language; they designed conceptual representations, built systems that analyzed natural language expressions into those representations, and they simulated human memory and common sense inference in systems that performed tasks such as narrative comprehension, question answering, paraphrasing, and summarization (Quillian 1968; Schank 1975; Anderson and Bower 1973).

One of the first demonstrations of *conceptual analysis* was a system called MARGIE (Schank et al. 1975), which analyzed natural language input and generated natural language output in the form of inferences and paraphrases. To analyze its input text, MARGIE built a representation of its meaning in a system called Conceptual Dependency (Schank 1972; 1975), a semantic primitive system (Schank 1972; Wilks and Fass 1992; Jackendoff 1983; Wierzbicka 1996; Wilks 1996; Winograd 1978) whose structures reflect the pictures and representations in people's minds about real-world acts and events, as well as the changes in the state of the world that result from them. The theory behind Conceptual Dependency proposes that the conceptual structures that a human understander of language manipulates are not isomorphic to words, phrases, or grammar of their language, but that the thought process behind language understanding takes place in a "private" realm of "language of thought," having different origins and, therefore, different characteristics from the spoken language of a human understander (Miller and Johnson-Laird 1976; Fodor 1975; Schank 1975). Conceptual Dependency (CD) decomposes meanings into complex structures based on combinations of "language-free" conceptual primitives, comprising a thought representation of the actual events separate from the language. Figure 1 shows examples of sentences and their conceptual analyses in CD. At the same time, building these systems by hand incurs the high cost of knowledge engineering (Feigenbaum 1977) of the symbolic structures that comprise the conceptual analysis.

Crowdsourcing is now well known as an inexpensive and fast method to collect human annotation of datasets for classifiers of natural language based on machine learning. When researchers have approached the viability of crowdsourcing for annotations of natural language corpora, the main question has been about whether people recruited from crowds can provide performance and annotation quality that is comparable to experts (Callison-Burch 2009; Snow et al. 2008). Crowdsourcing has been used to collect commonsense knowledge, narrative intelligence knowledge, and annotations of FrameNet (Havasi, Speer, and Alonso 2007; Li et al. 2012; Boujarwah, Abowd, and Arriaga 2012; Chang et al. 2015). In all of these cases, however, the annotations and knowledge structures that crowdworkers provide are in the form of natural language. While studies have shown that crowdworkers can grasp language-free primitives in coherent ways to collect commonsense knowledge (Johnson-Laird and Quinn 1976; Macbeth and Barionnette 2016), no research to date has addressed the challenges of using crowdsourcing to collect complex structures based on these language-free primitives to support conceptual analysis systems.

Research Questions

We designed a study with crowdworkers as human participants to determine whether we could build a parallel corpus for conceptual analysis systems through crowdsourcing. We wanted to know if crowdworkers could be presented with natural language expressions—in our case, simple English sentences—and could they, through our guided process, translate their conceptualizations of the meanings of the sentences into a non-linguistic, primitive decomposition-based conceptual representation. We were interested in answering the following research questions (RQs) in our investigation:

- **RQ1: Capability.** Can crowdworkers build simple conceptualizations in a language-free conceptual base like Conceptual Dependency?
- **RQ2: Complexity.** Can crowdworkers build complex or

compound conceptualizations by connecting simple conceptualizations through connectives?

- **RQ3: Quality.** Do crowdworkers build conceptualizations reliably that agree strongly with a gold standard generated by experts?
- **RQ4: Refinement.** Can subsequent rounds of crowdsourcing be used to filter out the best conceptualizations built by earlier rounds of crowdsourcing?

First Study

To address these questions, our first empirical study presented participants with simple English sentences, and asked them to build language-free conceptualizations corresponding to the acts in those sentences.

CD Primitives and the Instrument Case

We chose the Conceptual Dependency (CD) meaning representation for our study because it intends to be a nonlinguistic conceptual representation, because it has relatively few primitives, and because it was applied successfully to natural language understanding systems over several decades (Schank and Abelson 1977; Schank and Riesbeck 1982; Lytinen 1992). CD represents meaning by decomposing it into complex combinations of primitive states, events, and acts. Although CD is usually presented as having 11 primitive acts and several methods for connecting primitive acts, here we focused on collecting conceptualizations built only from CD's six physical primitives, and its "instrument" case, which forms a connection between two primitive acts. The CD physical primitives are[1]:

PROPEL: To apply a force to.

MOVE: To move a body part.

INGEST: To take something to the inside of an animate object.

EXPEL: To take something from inside an animate object and force it out.

GRASP: To physically grasp an object.

PTRANS: To change the location of something.

We define a *primitive conceptualization* or *simple conceptualization* of an act to be one that specifies the primitive act, the actor, the object, and, occasionally, a direction case. A primitive conceptualization may also have an *instrument* case as an arc to another primitive conceptualization: This represents that the actor performs one primitive act as a part of accomplishing another primitive act. The instrument case is frequently used in CD representations of physical action verbs, where the conceptual structure consists of two or more physical primitive act structures with the same actor and with certain acts being the instruments of others. We call a larger conceptualization composed of more than one primitive conceptualization in this way a *complex conceptualization*.

Sentences and Corresponding CD Diagrams

For our study we created six simple English sentences. We constructed sentences for which we thought participants would conceptualize more than one CD primitive act and would see one act as being the instrument of the other act. We also constructed six gold standard CD conceptual structures corresponding to these sentences where each had two primitive acts, with one serving as the instrument for the other. The sentences and their corresponding gold standard CD diagrams were inspired by those in Schank (1975) and are shown in Figure 1.

For example, for the sentence "Lisa kicked the ball," the diagram is a PROPEL primitive conceptualization with Lisa as the actor and the ball as the object (Lisa PROPELing the ball). The PROPEL has a MOVE primitive conceptualization as its instrument, with Lisa as the actor, Lisa's foot as the object, and the ball as the "to" direction case (Lisa MOVEing her foot toward the ball).

Conceptualization Questionnaire

To study crowdworkers creating CD structures corresponding to the sentences, we presented participants with descriptions of the six CD physical primitives, and presented the sentence of interest as the *target sentence*. The descriptions of the CD physical primitives did not include the traditional names of the primitives given above, since we did not want participants to be biased by the names; instead each primitive was identified with a number.

We then asked a number of questions to crowdworkers about how they conceptualized the meaning of the target sentence. The questionnaire had three parts. The first part asked questions about the details of the *primary primitive act* in the target sentence to determine the physical primitive, the actor, the object, and any directionality to the act. The second part of the questionnaire asked questions about a possible *secondary primitive act* by asking, "Did the actor do anything else as part of the target sentence?" and by asking if there was a second action corresponding to one of the other primitives that "made the first action happen." Following this was the third part to the questionnaire which consisted of a series of six questions which were identical to those in the first part, except that they focused on the secondary primitive act instead of the primary primitive act.

The questionnaire also asked for brief explanations of the choices of primitive acts, and provided a dynamic element showing a natural language gloss description of the conceptualization based on the users' entries in the form. Overall, the questions corresponded to the parts of the gold standard complex conceptualizations which were composed of two simple, primitive conceptualizations connected by an "instrumental" link. The HTML interface to the questionnaire is shown in Figure 2.

We conducted the study on Amazon Mechanical Turk by creating six human intelligence tasks (HITs) of our questionnaire, one for each target sentence. We estimated the time to complete the task at 15 minutes, and Turkers were rewarded $1 for completing the task. We accepted all 20 submissions for each target sentence from 20 unique Mechanical Turk

[1] These brief descriptions are from Schank (1975).

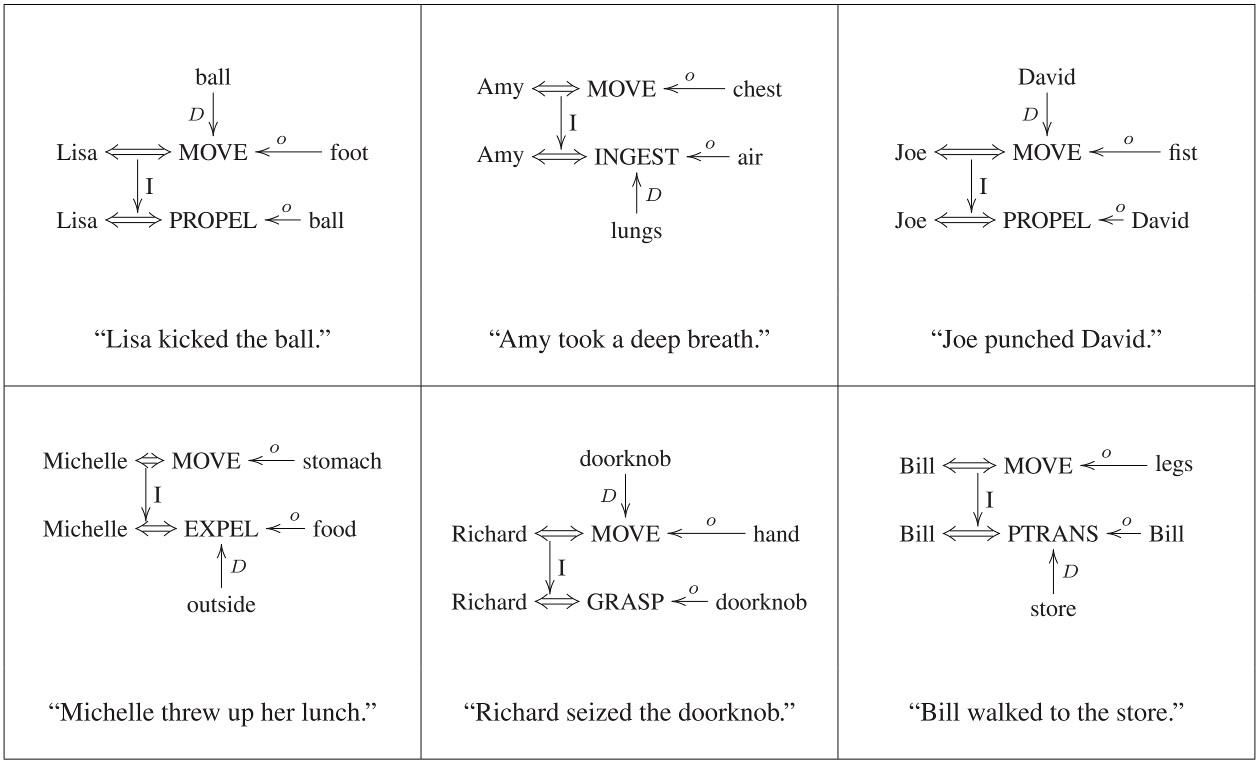


Figure 1: Sentences and corresponding gold standard Conceptual Dependency diagrams. Double arrows point to the actor, and single arrows marked “o”, “D”, and “I” indicate the object, the “to” directional case, and the instrument case, respectively.

workers, for a total of 120 submissions, and no submissions were rejected. To qualify to perform the task, the Turkers were required to be “Masters” workers with at least 1000 HITs approved. Turkers used an average time of 10 minutes and 45 seconds and a median time of 6 minutes and 40 seconds to complete the task.

First Study Results

Building Simple Primitive Conceptualizations

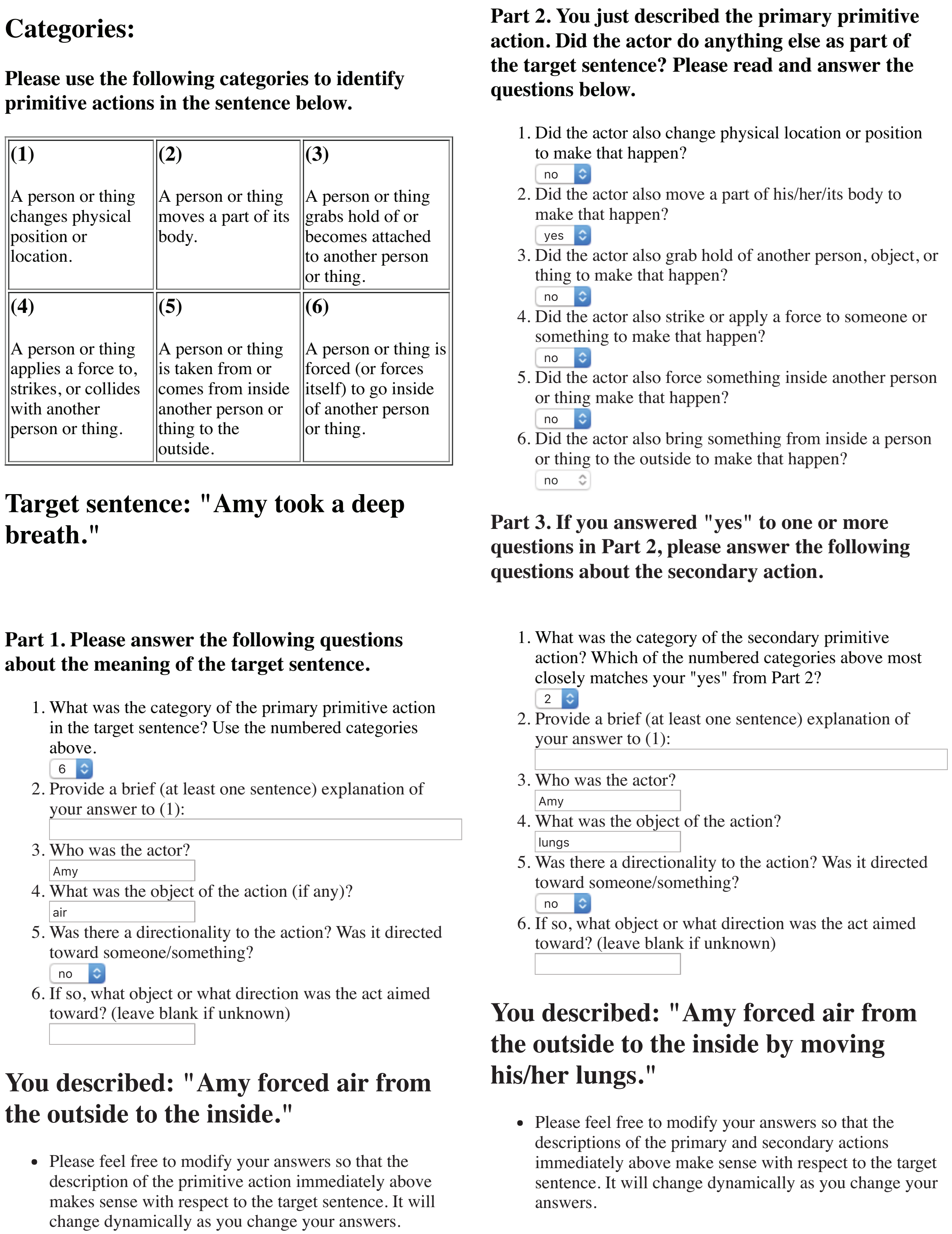
For our RQ1, we wanted to know if participants were at least able to build simple conceptualizations in CD, a language-free conceptual base, through the questionnaire. Simple primitive conceptualizations consisted of a selection of a primitive act, specification of an actor and object, and possibly specification of a direction for the act. For the six sentences, all submissions by participants provided at least a primary primitive act and an actor. 115 out of 120 submissions by participants (96%) included an object for the primary primitive conceptualization.

Many submissions provided both primary and secondary primitive conceptualizations. For example, for “Lisa kicked the ball”, 20 participants provided 40 primitive conceptualizations in total. The most common two primitives provided were PROPEL and MOVE, which matched primitives in the gold standard CD structure. As a second example, for “Amy took a deep breath,” 20 participants provided 28 primitive conceptualizations. In that case, the most common two primitives provided were MOVE and INGEST, which matched primitives in the gold standard CD structure. Table 1 gives examples of the object and direction cases for the two most common primitive conceptualizations for the “Lisa” and “Amy” sentences.

Building Complex Conceptualizations

We wanted to know if participants were able to build complex conceptualizations that connected simple primitive conceptualizations through connectives like instrumentation (RQ2). Participants built complex conceptualizations corresponding to the sentences through the questionnaire. In Part 1, they provided a “primary primitive act” conceptualization, and, in Part 2, they provided a “secondary primitive act” that served as the instrument case for the primary primitive act. In each case they were required to choose one of the six physical primitives as corresponding to the action.

Nearly all participants provided a primary conceptualization, but the questionnaire gave participants the option of providing a second conceptualization or not providing it. In a majority of cases, participants did provide a complex conceptualization corresponding to the sentence. In fact, 92 out of 120 (77%) of the submissions over the six sentences were complex conceptualizations. For “Lisa kicked the ball”, all 20 of the conceptualizations provided were complex and consisted of a pair of sub-conceptualizations. For “Amy took



Categories:

Please use the following categories to identify primitive actions in the sentence below.

(1) A person or thing changes physical position or location.	**(2)** A person or thing moves a part of its body.	**(3)** A person or thing grabs hold of or becomes attached to another person or thing.
(4) A person or thing applies a force to, strikes, or collides with another person or thing.	**(5)** A person or thing is taken from or comes from inside another person or thing to the outside.	**(6)** A person or thing is forced (or forces itself) to go inside of another person or thing.

Target sentence: "Amy took a deep breath."

Part 1. Please answer the following questions about the meaning of the target sentence.

1. What was the category of the primary primitive action in the target sentence? Use the numbered categories above.
   6
2. Provide a brief (at least one sentence) explanation of your answer to (1):
3. Who was the actor?
   Amy
4. What was the object of the action (if any)?
   air
5. Was there a directionality to the action? Was it directed toward someone/something?
   no
6. If so, what object or what direction was the act aimed toward? (leave blank if unknown)

You described: "Amy forced air from the outside to the inside."

- Please feel free to modify your answers so that the description of the primitive action immediately above makes sense with respect to the target sentence. It will change dynamically as you change your answers.

Part 2. You just described the primary primitive action. Did the actor do anything else as part of the target sentence? Please read and answer the questions below.

1. Did the actor also change physical location or position to make that happen?
   no
2. Did the actor also move a part of his/her/its body to make that happen?
   yes
3. Did the actor also grab hold of another person, object, or thing to make that happen?
   no
4. Did the actor also strike or apply a force to someone or something to make that happen?
   no
5. Did the actor also force something inside another person or thing make that happen?
   no
6. Did the actor also bring something from inside a person or thing to the outside to make that happen?
   no

Part 3. If you answered "yes" to one or more questions in Part 2, please answer the following questions about the secondary action.

1. What was the category of the secondary primitive action? Which of the numbered categories above most closely matches your "yes" from Part 2?
   2
2. Provide a brief (at least one sentence) explanation of your answer to (1):
3. Who was the actor?
   Amy
4. What was the object of the action?
   lungs
5. Was there a directionality to the action? Was it directed toward someone/something?
   no
6. If so, what object or what direction was the act aimed toward? (leave blank if unknown)

You described: "Amy forced air from the outside to the inside by moving his/her lungs."

- Please feel free to modify your answers so that the descriptions of the primary and secondary actions immediately above make sense with respect to the target sentence. It will change dynamically as you change your answers.

Figure 2: The HTML user interface for the crowdsourcing task for the first study, showing the descriptions of the conceptual primitives as categories, and typical answers for the target sentence "Amy took a deep breath." The "You Described" sections of the interface provide a natural language gloss description of the conceptualization based on the users' entries in the form. These elements are scripted to dynamically update when the user updates their entries.

Target Sentence	CD Primitive Act	Object	Direction
“Lisa kicked the ball.”	MOVE	“foot” “leg” “ball”	“ball”
	PROPEL	“ball”	“ball”
“Amy took a deep breath.”	MOVE	“diaphragm” “respiratory system” “chest” “lungs” “process of inhaling”	N/A
	INGEST	“air” “breath” “oxygen”	“inward” “her lungs” “Amy”

Table 1: Examples of object and direction cases for the most common primitives provided by participants for two of the target sentences. Results are arranged by target sentence to show the two most common primitives for each sentence. No participant provided a direction case for a MOVE primitive for the sentence “Amy took a deep breath.”

a deep breath”, however, only eight out of 20 conceptualizations were complex. In summary, the majority of conceptualizations crowdworkers provided were complex, but crowdworkers’ willingness to provide a complex conceptualization was strongly dependent on the target sentence.

Quality of Conceptualizations

Regardless of whether or not crowdworkers can provide us with complex conceptualizations, there are key questions regarding the quality of the conceptualizations they provide (RQ3). We wanted to know if the conceptualizations they provided agreed with the gold standard conceptualization for that target sentence (shown in Figure 1). On the basis of which primitive conceptualizations the participants provided, for the target sentence “Lisa kicked the ball,” 16 out of 20 participants created a conceptualization with PROPEL as the primary primitive and MOVE as the secondary primitive (MOVE as the instrument of a PROPEL), which agreed with the gold standard. Three more participants had PROPEL as primary and MOVE as secondary (PROPEL as an instrument of a MOVE), which only disagreed with the gold standard in the sense of the direction of the instrumental link. When the orientation of the instrumental link is disregarded, participants agreed with us and themselves in 19 out of 20, or 95%, of cases for the “Lisa” sentence.

However, for the target sentence “Amy took a deep breath”, only 4 out of 20 conceptualizations resembled the gold standard expectations of Amy INGESTing air with Amy MOVEing her chest as an instrument case. The conceptualizations were dominated by a simple MOVE primitive conceptualization, which was given by eight participants.

For each conceptualization that was submitted, we calculated a score representing how closely the structure matched our gold standard structure for that sentence, which we called the *gold standard match score*. For each simple conceptualization structure there were four items: the actor case, the primitive type, the objective case, and the direction case. In participants’ submissions, these cases were provided as free-form text, so, in order to accurately determine the agreement between the gold standard and the participants’ submissions, we removed articles, possessives, and possessive pronouns like “the,” “her,” and “Lisa’s” from the text answers to these questions. When both primitive types of a complex conceptualization were correct, we also counted the direction of the instrumental link as a factor in the matching score.

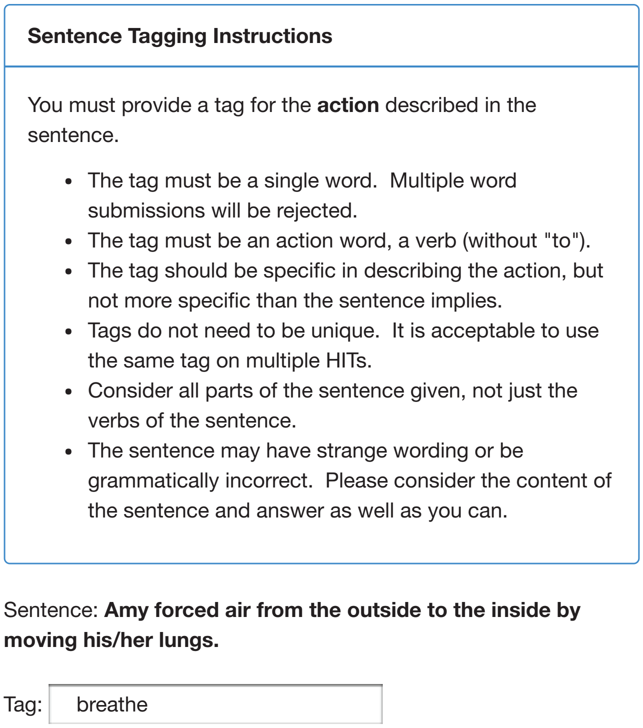
Overall, participants submissions had an average gold standard match score of 65%. Not all of the disagreement could be attributed to lack of a secondary primitive act for many submissions; when the agreement measurement was constrained to only submissions that were complex conceptualizations, the agreement with the gold standard increased only modestly to 74%. Clearly some of the disagreement was due to slightly different ways of referring to objects: For example, “diaphragm”, “chest”, “lungs”, and “respiratory system” all seem to refer to the same part of the body for the MOVE primitive act involved in “Amy took a deep breath” (Table 1). But generally, this underscores the need for more sophisticated systems for singling out the best conceptual structures created and submitted by crowdworkers.

Second Study

In our second study, we investigated whether subsequent rounds of crowdsourcing could be used to select only the best conceptualizations for building and refining a parallel corpus for conceptual analysis (RQ4). We developed a method for refining the conceptualizations that presents them to a second set of crowdworkers who attempt to “recompose” the action word from the original sentence. This sort of quality control step is common in studies of crowdsourcing knowledge for intelligent systems (Li et al. 2012; Boujarwah, Abowd, and Arriaga 2012).

Presenting Complex Conceptualizations

To present the conceptual structures to crowdworkers for refinement, we employed a simple natural language generation scheme, converting the structures into English sentence



Sentence Tagging Instructions

You must provide a tag for the **action** described in the sentence.

- The tag must be a single word. Multiple word submissions will be rejected.
- The tag must be an action word, a verb (without "to").
- The tag should be specific in describing the action, but not more specific than the sentence implies.
- Tags do not need to be unique. It is acceptable to use the same tag on multiple HITs.
- Consider all parts of the sentence given, not just the verbs of the sentence.
- The sentence may have strange wording or be grammatically incorrect. Please consider the content of the sentence and answer as well as you can.

Sentence: **Amy forced air from the outside to the inside by moving his/her lungs.**

Tag: breathe

Figure 3: The HTML user interface for the "recomposition" sentence-tagging task.

glosses for presentation to the crowdworkers. The gloss of the decomposition was generated systematically by creating a sentence with two clauses corresponding to the two primitive conceptualizations that combined to form the complex conceptualization. The main clause of the sentence corresponded to the primary primitive act of the sentence, and was connected to the secondary primitive act of the sentence using "by", indicating its instrumental role. Each clause had the main actor (subject) and object given in the conceptualization, and used a verb corresponding to the conceptual primitive. When a direction was given, a prepositional phrase was added to the clause with the preposition "toward" and the direction object. The second clause, beginning with "by", was added with a gerund form of the verb corresponding to the primitive and object and direction object prepositional phrase. These glosses were identical to those shown in the "You Described" feedback in the interface to the first study. In this second study, we presented only sentences generated from the 92 complex conceptualizations that were generated in the first study.

Recomposing the Action Word

We presented crowdworkers with the generated sentences and asked them to provide an action word, (a verb in infinitive form without "to") as a tag for the action described in each sentence. The interface to this task is shown in Figure 3. This second study was also performed on Mechanical Turk, where participants were rewarded $0.20 per HIT. We created 92 tagging tasks based on the sentences generated from the complex conceptualizations in our first study, and we allowed 10 unique workers to work on each, for a total of 920 assignments. Participants were required to have a HIT approval rate greater than 90%, and eight submissions were rejected because they were not single action words. Participants took an average time of 69 seconds and a median time of 25 seconds on the assignments.

After we stemmed and lemmatized the responses, we scored the generated sentences according to the number of responses that matched the action verb from the original sentence in the first study (or a close synonym of it). Figure 4 shows the top scoring complex conceptual structures built by crowdworkers for the "Lisa" and "Amy" sentences. This provided us with 92 observations to use to calculate a statistical measure of the correlation between the recomposition match score, which ranged from 0 to 10, and the gold standard match score, which ranged from 0 to 9. A calculation of Spearman's ρ came out to 0.409, and it was a statistically significant correlation ($p < 0.001$). Overall, when limiting the corpus from our first study to the subset of submissions that had the highest recomposition score for each original sentence, the average gold standard match score increased to 85%.

Discussion

Examining several different target sentences provided the opportunity to compare and contrast a variety of performances by crowdworkers on the task. Looking at the frequency with which crowdworkers created complex conceptualizations that agreed with our gold standard, and the degree to which crowdworker's complex conceptualizations agreed with each other, sentences like "Lisa kicked the ball" were far more successful than other sentences, such as "Amy took a deep breath", and "Michelle threw up her lunch".

On further examination, we realized one reason for the discrepancy may have been that the description of the INGEST primitive that we provided to crowdworkers, which read, "A person, object, or thing is forced (or forces itself) to go inside of another person, object, or thing", did not account well for the possibility of INGESTing air or another gas. More generally, we recognized that our attempts to crowdsource conceptualizations for a corpus will provide an opportunity to evaluate the suitability of the conceptual representation system itself. Earlier work on conceptual primitive systems in artificial intelligence circles rarely sought to determine whether the set of primitives were psychologically valid, whether they could be improved, or whether they could even be reconceived from scratch using data gathered in human-subjects experiments.

Since our motivation was the desire to reduce the time and costs of collecting such a parallel corpus, we consider the issue of the cost of using this technique. Because the recomposition task simply involves reading a sentence and providing a single-word tag—data that likely can be obtained from the crowd for fractions of a cent, or even for free as a qualification task—we focus on the cost of gathering the CD conceptualizations.

For each sentence in our study we accepted 20 CD diagrams as submissions, each of which cost $1. However, we

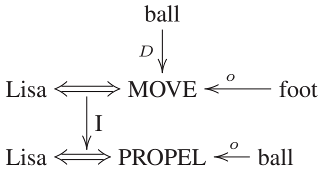
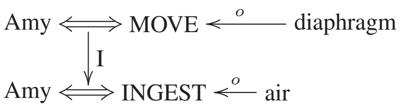
Crowdsourced CD Diagram	Generated Sentence	Action Word and Recomposition Score	Gold Standard Match Score
 ball D Lisa MOVE o foot I Lisa PROPEL o ball	“Lisa struck or applied a force to the ball by moving his/her foot toward the ball.”	“kick” Score: 9/10	Score: 9/9
 Amy MOVE o diaphragm I Amy INGEST o air	“Amy forced air from the outside to the inside by moving his/her diaphragm.”	“breathe”, “inhale”, “exhale” Score: 8/10	Score: 7/9

Figure 4: The top scoring crowdsourced CD diagrams built by “decomposition” crowdworkers for two of the six original sentences (“Lisa kicked the ball”, and “Amy took a deep breath”) along with the generated sentence which was presented to the “recomposition” workers, the expected action word(s), the recomposition score, and gold standard matching score of the CD structure.

may not need 20 CD diagram submissions to receive one with a high gold standard matching score. We found that of the 120 CD diagrams received in our first study, 52 out of 120, or 43%, had gold standard matching scores of 8 or 9, where 9 was the highest score possible. If p represents this 43% probability, we can calculate the probability of receiving at least one high-scoring diagram in a set of n trials to be $1 - (1 - p)^n$. If we set n to five trials (i.e., five CD diagram submissions), there is a 94% probability of receiving at least one high-scoring diagram. At our experimental reward rate, this would cost $5.00 per sentence, but we note that, in designing the tasks for this study, we focused simply on whether crowdsourcing was possible in this domain, and did not seek out lower per-assignment rewards.

Conclusion and Future Work

In this investigation we set out to show that crowdsourcing may provide leverage against the task of collecting a parallel corpus for building conceptual analysis systems that can generate a language-free representation of the meaning of a text. Our empirical studies demonstrated that crowdworkers could provide us with a variety of both simple and complex conceptualizations. We found a statistically significant correlation between the quality of the conceptualizations produced by the crowd, and the fraction of crowdworkers who could figure out the original lexical item with only the decomposed conceptualization as input. This proves that using multiple phases of crowdsourcing with our recomposition method can be effective way to defeat the noise typically present in crowdsourcing platforms, and encourages future work to build a parallel corpus matching natural language to high-quality non-linguistic conceptual structures. The studies described in this paper were limited to crowdsourcing the conversion of entire sentences into a decomposed conceptual structures. However, our eventual goal is a conceptual analysis system, built using machine learning methods, that automates this conversion.

Our current study only examined the six physical primitives of Conceptual Dependency and the instrumentation connective with a corpus of only six sentences. For a future evaluation and validation with greater generalizability, we can use a larger and more realistic test corpus of English (Ide and Macleod 2001) and explore techniques for increasing efficiency and reducing costs of crowdsourcing in comparisons with expert performance. Future studies will consider other CD primitives for mental and social acts (such as MTRANS), and connecting primitive acts through causation, or they may also consider conceptual representation frameworks different from CD altogether. Finally, crowdsourcing for even larger conceptualizations, such as those in scripts (Li et al. 2012; Schank and Abelson 1977), has yet to be attempted in a language-free conceptual base like CD. Future investigations of these issues will bring us closer to scaling up conceptual representation systems for computers to achieve human-like conceptual understanding of natural language, and, importantly, they will reduce the time and cost to build them.

Acknowledgments

We thank the anonymous reviewers for their thoughtful and insightful comments about our paper, and the organizers for providing us with an opportunity to revise our paper.

References

Anderson, J. R., and Bower, G. H. 1973. *Human Associative Memory*. Washington, DC: Winston.

Boujarwah, F.; Abowd, G.; and Arriaga, R. 2012. Socially computed scripts to support social problem solving skills. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1987–1996. ACM.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286–295. Association for Computational Linguistics.

Chang, N.; Paritosh, P.; Huynh, D.; and Baker, C. 2015. Scaling semantic frame annotation. In *The 9th Linguistic Annotation Workshop, held in conjuncion with NAACL-HLT*, 1–10.

Feigenbaum, E. A. 1977. The art of artificial intelligence. 1. themes and case studies of knowledge engineering. Technical Report STAN-CS-77-621, Stanford University, Department of Computer Science, Stanford, CA.

Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Havasi, C.; Speer, R.; and Alonso, J. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*.

Hirschberg, J., and Manning, C. D. 2015. Advances in natural language processing. *Science* 349(6245):261–266.

Howe, J. 2006. The rise of crowdsourcing. *Wired Magazine* 14(6):1–4.

Ide, N., and Macleod, C. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics*, volume 3, 1–7. Lancaster University Centre for Computer Corpus Research on Language Lancaster, UK.

Jackendoff, R. S. 1983. *Semantics and Cognition*, volume 8 of *Current Studies in Linguistics*. Cambridge, MA: MIT Press.

Johnson-Laird, P., and Quinn, J. 1976. To define true meaning. *Nature* 264:635–636.

Johnson-Laird, P.; Robins, C.; and Velicogna, L. 1974. Memory for words. *Nature* 251:704–705.

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Li, B.; Appling, D. S.; Lee-Urban, S.; and Riedl, M. O. 2012. Crowdsourcing narrative intelligence. *Advances in Cognitive Systems* 2:25–42.

Lytinen, S. L. 1992. Conceptual dependency and its descendants. *Computers & Mathematics with Applications* 23(2):51–73.

Macbeth, J. C., and Barionnette, M. 2016. The coherence of conceptual primitives. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*. the Cognitive Systems Foundation.

Miller, G. A., and Johnson-Laird, P. N. 1976. *Language and Perception*. Cambridge, MA: Belknap Press.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Paivio, A. 1971. *Imagery and verbal processes.* New York, NY: Holt, Rinehart & Winston.

Quillian, M. R. 1968. Semantic memory. In Minsky, M., ed., *Semantic Information Processing*. Cambridge, MA: MIT Press. 227–270.

Rosch, E. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104(3):192.

Sachs, J. 1967. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics* 2(9):437–442.

Schank, R. C., and Abelson, R. P. 1977. *Scripts, Plans, Goals and Understanding : An Inquiry Into Human Knowledge Structures*. Hillsdale, N.J.: L. Erlbaum Associates.

Schank, R. C., and Riesbeck, C. K. 1982. *Inside Computer Understanding: Five Programs Plus Miniatures*. Hillsdale, NJ: L. Erlbaum Associates Inc.

Schank, R. C.; Goldman, N. M.; Rieger III, C. J.; and Riesbeck, C. K. 1975. Inference and paraphrase by computer. *Journal of the ACM (JACM)* 22(3):309–328.

Schank, R. C. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology* 3(4):552–631.

Schank, R. C. 1975. *Conceptual Information Processing*. New York, NY: Elsevier.

Schuler, K. K. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263. Association for Computational Linguistics.

Wason, P. C., and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content*. Harvard University Press.

Wierzbicka, A. 1996. *Semantics: Primes and Universals*. Oxford University Press, UK.

Wilks, Y., and Fass, D. 1992. The preference semantics family. *Computers & Mathematics with Applications* 23(2-5):205–221.

Wilks, Y. 1996. Good and bad arguments for semantic primitives. Technical report, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico,.

Winograd, T. 1978. On primitives, prototypes, and other semantic anomalies. In *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '78, 25–32. Stroudsburg, PA, USA: Association for Computational Linguistics.