# A Kids' Open Mind Common Sense*

**Antal van den Bosch** and **Pim Nauts** and **Nienke Eckhardt**
Tilburg center for Cognition and Communication
Tilburg University
Antal.vdnBosch@uvt.nl

**Abstract**

We propose a collaborative approach to the issue of resource creation for commonsense computing by developing a collaboratory application aimed at children. Human validation is enabled through a game-with-a-purpose (GWAP) interface, gathering reliability judgements of assertions that can be used to aid the process of resource validation. Our experiments confirm that children aged 10 to 12 can be valuable and reliable partners in building commonsense databases, due to their stage of mental development and their eagerness to play GWAPs. Results show that children adapt their word choice in the assertions they provide to the difficulty level of the stimuli words, and that the judgements gathered through in-game validation can help to validate about 30% of the gathered statements automatically.

## Introduction

Endowing machines with the ability to reason over human common sense has been argued to be a vital step towards computers achieving a truly supportive role in everyday life; assisting humans with everyday tasks and needs (Minsky 2006). However, common sense is a field of great scale, and to date, capturing it for computational use is still surrounded by unsolved problems that relate to the very nature of human knowledge and language (Singh 2003).

If we want to capture and harvest human common sense for further processing by computers, a first step is to somehow explicate that common sense. People are known to be a troublesome source for explicating commonsense knowledge. Each person's common sense is acquired largely subconsciously throughout a lifetime, and is mostly taken for granted once learned. Paradoxically, we tend to be blind to the intricateness of the knowledge we have acquired, especially commonsense knowledge (Minsky 2006). Yet over the years we have built up an impressive mental commonsense repository. Instead of probing adults, we could try to tap into the minds that are most active in internalizing it: those of children.

*This work is supported by the Netherlands Organisation for Scientific Research (NWO) as part of the NWO Vici "Implicit Linguistics" project. The authors wish to thank Peter Berck, Sander Wubben, and Marga van Zundert for their support.


To children, acquiring and applying common sense is an important part of their everyday learning and reasoning. Children are explicitly expanding their mental databases every day and seem to have an innate curiosity for all facts of life that appear so obvious to adults. Some of this natural curiosity may be harnessed for building a commonsense resource. However, while turning to children could possibly solve the implicitness problem, it creates two types of other problems that relate to the nature of childhood. First, children are naive in their views on the world; many of their assertions may be incomplete or false. We may not safely assume children will provide us with truthful commonsense knowledge. Therefore, we need ways to account for possible problems with truthfulness. Second, children cannot simply be coerced in performing any task, even if it touches their innate interest in learning about the world; we need to develop a method to motivate children to help us.

Concerning the scale of the problem, there has been a recent trend in AI research to use collaborative approaches (or *human computing*) in solving large-scale problems at which humans still outperform computers (Stork 1999), e.g. in adding game elements (Von Ahn, Kedia, and Blum 2006; Von Ahn and Dabbish 2008) to harvest valuable information generated as a side-effect of gameplay. If we can devise a strategy to use the natural curiosity of children and their willingness to play, we may at least have a partial solution to the issue of scale.

We have developed a GWAP-based collaborative approach to commonsense resource development, aimed at children. The GWAP is two-phased. In the first phase of the game children are given a concept to describe (e.g. "boat"), for which they can choose from a limited number of templates (e.g. "isA"). In the second phase, to provide a type of in-game human validation typical for GWAPs, children are asked to guess the given concept based on another user's assertion with the subject masked (e.g. "... is a means of transport."). Issues that might arise due to the involvement of children (truthfulness, egocentric perspectives, general knowledge levels) can be partly countered through targeting certain age groups, keeping the possibility open that it also appeals to and can be played by all ages above the target age group as well.

Because the majority of previous research has focused on adults, we need to establish a new link with work on cog-

nitive development in children. First, we briefly review related work in both areas. Subsequently, we describe the experiments we carried out with our two-phased GWAP and children participants. After providing the results of our experiments, we discuss our findings, and state our conclusions and recommendations.

## Related Work

Within the field of artificial intelligence, acquiring commonsense knowledge has been recognized as a bottleneck problem since the early 1960s (Kuipers 2004), and has been a long-standing topic of research (McCarthy 1959; McCarthy and Hayes 1969; Singh 2002; Minsky 2006; Speer et al. 2009). It has led to the development of a number of systems centering around commonsense resource creation, such as Cyc, MindPixel, ThoughtTreasure, and the Open Mind Common Sense (OMCS) initiative, part of the collaborative Open Mind Initiative (Stork 1999). The Open Mind Initiative (OMI) explicitly utilizes the growing number of nonspecialist users ("netizens") found online, with the Web as a low-cost framework for collecting data.

The data OMCS gathers is added to ConceptNet, a semantic network in which concepts are related through a limited set of predicate types (Liu and Singh 2004). Instantiations of predicates are stored in ConceptNet as triples connecting subjects and objects to predicate verbs. OMCS and ConceptNet together form a notable effort to counter the problem of harvesting commonsense knowledge at a realistic scale. Nonetheless, the question arises whether the approach of OMCS is the most effective possible: it is dependent on the willingness of contributors; users donating their time to a project that offers arguably little entertainment value.

Our solution to collaborative commonsense resource development is derived from an idea proposed by Von Ahn and Dabbish (2008), *Games With A Purpose* (GWAPs). By using the constructive channeling of human intelligence through computer games, computationally hard problems that require human judgement or annotation can be solved, thereby helping to improve AI algorithms (Von Ahn and Dabbish 2008). The GWAP framework allows to collect data through a potentially large human contributor pool, regardless of the participants' interest in contributing to AI.

### Verbosity

One of the first and best-known examples of a GWAP is the ESP Game, in which players have to agree in labeling images. The information harvested in the game provides valuable metadata about what the image depicts (Von Ahn and Dabbish 2004). The ESP game cleverly allows for validating contributions as an integral part of the game. If paired players independently agree on a label, it is assumed the named concept is indeed depicted. Von Ahn and colleagues then developed Verbosity (Von Ahn, Kedia, and Blum 2006), a game for collecting commonsense facts, in which the process of contributing common sense is partly hidden and a gameplay element is introduced. In the game, two players are randomly paired and assigned the roles of either Narrator or Guesser. The Narrator is shown a 'secret' word and needs to describe it in such a way that the Guesser can guess the word in turn. When the Guesser is able to name the concept given to the Narrator based on the given descriptions, these are confirmed as reliable in relation to the target concept and thus validated as a truthful piece of common sense.

### Commonsense development in children

In our aim to bridge the gap between research on commonsense knowledge acquisition, representation, and children's cognitive development, we connect what Marvin Minsky and Jean Piaget have stated on the two points.

In *The Society of Mind* Minsky states his views on cognitive developmental processes in children. According to Minsky and Papert, children learn from experiences and have the innate ability to give basic structure to these experiences which, over time, leads to logical capabilities (Minsky and Papert 1988). Another important concept Minsky adopts to explain the roots of the implicitness of commonsense knowledge, is that of *Infantile Amnesia*: humans are unable to recall events from their early childhood because we have not yet acquired the skills to remember them (Minsky 2006). Yet, children will have mastered a considerable amount of common sense knowledge by then.

Jean Piaget charted new areas of research in the cognitive development in children. His particular point of interest to our research is his focus on distinctive references at certain ages. Inhelder and Piaget identify four main stages in children's cognitive development: (i) Sensori-motor (ages 0–2), (ii) Pre-operational (ages 2–7), (iii) Concrete-operational (ages 7–11), and (iv) Formal-operational (ages 11 and onwards) (Inhelder and Piaget 1958). It is the *Formal-operational* stage that is of interest to our research, as it is the stage in which children learn in a systematic way to use symbols to relate to abstract concepts and to manipulate variables. Before this age, it may be possible to ask children to talk about concrete concepts, but less successfully about abstract concepts. Consequently, we aimed our GWAP and experiment to children of around the age of 11, which is also the age at which reading and writing skills have become reasonably automated.

## Experiments

To evaluate the proposed approach, we conducted two experiments using an online GWAP. Both experiments involve child participants aged 10 to 12, and are fully in Dutch.

### First experiment

**Participants** 123 children aged 10 to 12, 63 male and 60 female. Children were invited to play the game by sending their parents a letter, explaining some background, requesting the parents' consent and explicitly inviting them to go to the website with our game. Addresses were collected from the mailing list of the local Children's University, an initiative of Tilburg University and Eindhoven Technical University, The Netherlands[1] in which children in the last grades

[1] http://www.uvt.nl/kinderuniversiteit/

of primary school (i.e. up to age 12 in The Netherlands) follow lectures by university professors. No specific efforts to control the consistency of the group were made.

**Materials** We built an online game to collect the commonsense efforts of the children[2]. In the game a static fictional persona, Robot Rob, is used to both guide the children through the game and to embody the concept of 'computer'. Personae such as Robot Rob, visible in the right hand side of the screenshot provided in Figure 1, have proven to have positive effects on credibility, motivation and perception of experience (Ball et al. 1997; Van Mulken, André, and Muller 1998). Several human characteristics (a human name, an image of a humanoid robot, the use of child-directed natural language) were used to boost the character's effectiveness. The interface was built using a text-centered design with bright, attractive colors and as few attention-intrusive visual elements as possible. Input is contributed through standard HTML-based forms.

**Stimuli** In one game, subjects are given nine different concepts to describe according to three predicates to be chosen out of a list of six, totalling to 27 assertions generated per game. The assertion generation consists of (i) selecting a template, and (ii) completing the assertion by manually typing the third (object) part of the triple. We used the six most frequently occurring predicates in OMCS for our game (MadeOf, IsA, UsedFor, CapableOf, PartOf, and AtLocation), assuming that these would be the easiest to grasp for children aged 10–12. The six predicates used constitute 70% of all assertions in the English OMCS (Liu and Singh 2004).

Stimuli were selected from the "Woorden in het basisonderwijs" (*Words in primary school*) book and lexicon, developed for socio-linguistic research purposes by Schrooten and Vermeer in the early 1990s (Schrooten and Vermeer 1994). The lexicon consists of 26,590 lemmas collected from books used in primary education to children aged 4 to 12, and provides statistical indicators of which words can be expected to be acquired at a given age. All stimuli words are nouns, as nouns are the most likely subject of the six predicates. From all nouns we removed compounds, as Dutch has a productive compounding system and compound words can often be trivially described by their inherent compositionality ("a football IsA ball"). Subsequently, the remaining nouns were split into difficulty levels. The difficulty level of a word is calculated in Schrooten and Vermeer's lexicon by measuring the word frequency (how often a word appears in a corpus) and its spread (in how many texts a word can be found). The more often a word is presented to a child, the more likely the word is acquired. Less frequently occurring words are acquired at a later age (Schrooten and Vermeer 1994).

Based on a geometric mean of these two variables, Schrooten and Vermeer discretize the lexicon into subsets of words, where each subset is linked to grades in which the words should preferably be used in teaching. We further grouped the 26,590 lemmas into three aggregrate groups, labeled *easy* (familiar to children at age 6), *average* (familiar to children at age 9), and *difficult* (may already be familiar to children aged 10–12, but may not be acquired yet).

Next to the three difficulty levels, words were categorized as being abstract ("love") or concrete ("tree") according to the Cornetto lexical semantic resource, a WordNet for Dutch (Vossen et al. 2007). We make the distinction between abstract and concrete as it further divides the three difficulty levels in a meaningful way: "love" is *easy*, but since it is an abstract concept it is probably harder to describe than "kiss", a fellow but concrete *easy* word.

For each combination of difficulty level and the concrete/abstract distinction we selected 20 words, except for the *easy*–concrete group, for which we selected 40 words, as we noticed in a pilot test children were best encouraged by being presented with the occasional extra easy and concrete words. The selection was performed manually to avoid near-synonyms and strongly related terms to appear in the final selection. This gave us a total of 140 stimulus words.

**Gameplay** After a welcome screen, players were asked to fill out some info on their characteristics relevant to our research - name, age, gender, class (grade), and mother tongue. Instruction consisted of both a step-by-step visual explanation as well as text. Children received instructions before the start of a phase, and could also recall the information during play. All instructions were pre-tested in both a pilot study and by two pre-school teachers. The children were presented with nine concepts to describe, spread over the three difficulty levels and the abstract/concrete distinction. They could select one of six predicates from a drop down menu. Every predicate type could only be used once per concept. Assertions were entered in a text box, where the selected predicate is pre-filled by a templatic natural language phrase (such as "is gemaakt van", *is made of*, for MadeOf). For every concept three assertions had to be given. During every step it was made clear to the children that there was no such thing as a 'wrong' answer. They were also given the option to skip stimuli.

To further motivate the participating children we incorporated an element of competition. In the end, all players were rewarded with a second place in a top three as "second best teacher of Robot Rob" of the day, to encourage them to play it again and to not let them down—leaving the replacement of this working solution by a proper scoring system to future work.

Given the stage of development the participating children were in, we expected a certain eagerness to participate. The feedback given afterwards indicated this was very much the case. The persona that invited the children to help the computer was especially motivational, as it gave a sense of a collaborative effort.

## Second experiment

The second experiment, using a "guesser" GWAP, was aimed at determining the accuracy of the statements elicited

[2] http://ilk.uvt.nl/gezondverstand

in the first experiment. The experiment was carried out in a real-world setting, using questionnaires on paper instead of online, as we wanted to avoid look-ups and parental help. Note that our online GWAP does include the second game, which the player automatically enters after having played the first game. Figure 1 provides a screenshot of this second phase of the online GWAP.


Figure 1: Second phase of the game: guessing a concept based on assertions provided

**Participants** 119 10 to 12-year-old children in the last two grades of Dutch primary school (grades seven and eight), spread over four classes (two classes per grade). No specific efforts to control the consistency of the groups were made.

**Design** We used an independent measures design (between subjects). Correctness of the answers is used as the dependent variable, condition is used as the independent variable, measuring the effect of condition on correctness.

**Materials** The questionnaires consisted of an introduction to the guessing game, a brief example, and the key stimuli questions with supportive clues. Three versions were made, varying the number of clues provided. The stimuli were drawn directly from assertions stated online during the first experiment, given for 27 concepts selected from the 140 stimulus words of Experiment 1. One set of clues of one concept to be guessed is regarded as one stimulus. Children were asked to guess a concept based on stimuli.

**Stimuli** Each clue provided as part of a stimulus consists of an assertion in which the first (subject) noun is masked; this is the noun to be guessed. For example, a child user would see *. . . is an animal*, *. . . can be found near a farm*, and *. . . can moo*. Stimuli were presented per type (abstract vs concrete) and per difficulty level (easy, average, difficult). The concepts on which we based the stimuli were randomly sampled (using Simple Random Sampling with equal probabilities in PASW Statistics) for every difficulty level and type (abstract-concrete) using a split data file. We did not randomize between conditions; every child was presented the assertions for the same words since the proportion of our group of subjects was too small. The number of assertions (clues) per concept provided to the children was evenly varied between two, three, and four. Statistical analysis showed no significant effect of this number on the correctness of the answer.

**Procedures** The experiment was conducted in classrooms of participating primary schools. The questionnaires were filled out by all the children in a class simultaneously. The researcher first explained the general idea and children were then handed out the questionnaires. Questions could be asked and instructions were read aloud before the children were given 10 minutes to fill out the questionnaires. Children were not allowed to communicate during the experiment. All questionnaires were manually checked and scored by the researcher.

## Results

### First Experiment

Within three weeks after sending the letters to the children, our game had been played 150 times, excluding about 50 games played for demonstration purposes or with bad intent. The first genuine 150 games were played by 123 children; a subset of the children played it several times, with one child playing it over eight times. 63 Subjects were male, 60 were female.

| Assertion category | Number | Percentage |
|---|---|---|
| False | 13 | 1.3% |
| Poor | 32 | 3.1% |
| Correct, but typo | 37 | 3.6% |
| Correct | 943 | 92% |
| Total | 1,025 | 100.0% |

Table 1: Frequency of manually examined assertions, taken from a 25% random selection of the assertions.

A total of 4,077 assertions were collected. Table 1 provides a qualitative analysis of a 25% random selection of the 4,077 assertions into four categories of assertions. About 92% of the assertions were judged to be correct; an additional 3.6% could have been correct but were corrupted by typographical errors, which is not unexpected given that the children in the targeted age group are still learning how to spell words. Additionally, we observed a rather even distribution of provided assertions over the six OMCS templates; given a chance level of 16.7%, the strongest deviations are observed with the IsA template which was selected in 22.0% of all assertions, and the PartOf template, selected 12.8% of the time.

Figure 2 displays statistics on the relation in word usage in the provided assertions and the target words. Hypothetically, the difficulty level or the concreteness or abstractness of the target word may trigger the use of words in the same category in the assertions due to semantic relatedness of other

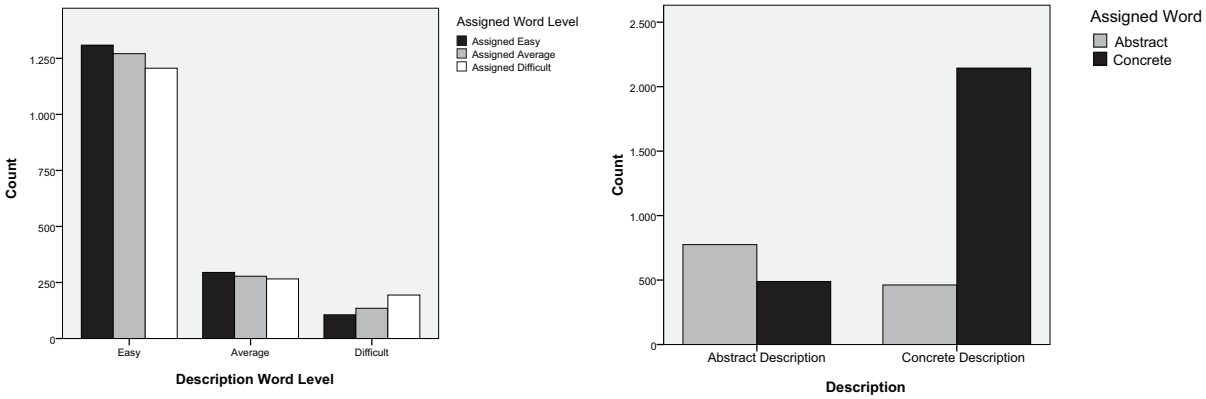


Figure 2: Relation between word usage in the provided descriptions (assertions) and the target words: difficulty level (left) and concreteness-abstractness distinction (right).

words in the triple, or to subconscious alignment. As the figure shows, there are indeed mild effects to be seen in difficulty level alignment (left) and concreteness or abstractness alignment (right). Difficult words were described with relatively more difficult words than easy words; abstract words were described with relatively more abstract words. The bars in the figure reflect actual numbers of words used in described assertions listed in the Schrooten and Vermeer lexicon (left) or in the Cornetto Wordnet (right).

### Second Experiment

For the second experiment, 119 questionnaires were filled out. Of a total of 2,142 stimuli, 658 or 30.7% were marked as correct. Table 2 provides a breakdown into types of errors. Roughly half of the guesses (51%) were incorrect or missing, while 22% of the guesses were semantically similar or related to the intended answers. Some guesses were strictly wrong, but were nonetheless reasonable alternate answers given the clues (22%). The concreteness or abstractness of the concept to be guessed influences correctness: 80.3% (529) of the 658 correct guesses are concrete nouns. The easiest group of concrete–*easy* nouns accounts for 38,7% of the correct guesses.

| Error type | Nr. of errors |
|---|---|
| Incorrect guess | 690 |
| Incorrect guess but possible given clues | 322 |
| Semantically related | 140 |
| Synonymous | 138 |
| No answer | 133 |
| Meronymous | 55 |
| Typos | 6 |

Table 2: Breakdown of the 1,484 incorrect guesses into error types

## Discussion

Our first experiment, with over 92% of the descriptions provided by the children being judged as reliable, shows that children aged ten to twelve are able to draw upon their common sense knowledge and their language skills to describe concrete and abstract words in the form of valuable assertions. When split on concreteness (93.4% correct) and abstractness (90.0% correct) a mild difference can be observed, which leads us to conclude that abstract words are not unreasonably more difficult to describe. As for differences in the age groups tested, 10-year-olds have described slightly more concrete and less abstract nouns than the 11 and 12-year-olds. Overall, also witnessed by the number of children that played the game several times, children report in after-experiment feedback time that they very much liked the task of describing words and providing the computer with common sense. The amount of reported fun surprised us, again revealing that the adult perspective on games like ours genuinely deviates from that of 10–12-year-olds.

The language used in the descriptions also yields some noteworthy results. Overall, most of the words the children used in their descriptions were easy and concrete, regardless of the concreteness or abstractness and difficulty level of the target concept. However, difficult target words were described more often with difficult words in the description; the same goes for assigned abstract words, which were described with relatively more abstract words.

The second experiment, aimed at the idea of in-game validation, indicates that guessing a concept based on the assertions provided by other children can contribute to collaborative validation of the descriptions, but the extent is relatively restricted. Of all the guessed words in the second experiment, 30.7% directly matched the target concepts, indicating that this portion of the data could be validated automatically using 100% string matching. An additional 22% of data can

in principle be validated by using more powerful (but error-prone) lookup in lexical semantic resources. In addition, our results indicate concrete nouns are easier to guess than are abstract nouns; this is in contrast with the relatively equal ease with which both categories are described.

Overall, our study provides indications that using a child-oriented GWAP can indeed lead to commonsense knowledge resource creation. We do have to state that our findings are only valid within the age group of our subjects (ten to twelve years of age); we did not run comparative tests with adults, adolescents or pre-teen children. Children before the age of 10 can be expected to lack some reading and spelling skills; children older than 12 may be less interested in playing than our focus group. While more research would be needed to confirm these expectations, we believe the age range 10–12 is the right target range for the type of GWAP we developed.

The main argument for conducting the second experiment in person instead of online was to rule out data obfuscation problems that might have occurred in the online game, for instance by asking for help from others. We should, however, note that the guessing task is harder than the assertion generation task, and that the number of validated assertions can be expected to be only a portion, currently 3 out of 10, of all assertions entered.

## Conclusion

It makes sense to adopt human computing to harvest the common sense knowledge we have all gathered and mastered throughout our lives. Realistically speaking, however, this scenario still has a long way to go until we arrive at the order of tens of millions of validated and trustworthy triples (Minsky 2006). In this paper we have proposed an alternative route to collecting commonsense facts, recruiting those among us who are right in the middle of acquiring it: children. We observed a considerable eagerness to play a Verbosity-like GWAP. The correctness of the assertions was quite high (92%), but we did observe that the guesser part of the game turned out quite a bit harder for children – only about 30% of the guesses matched the target word, allowing only this portion to be automatically validated.

If the proposed GWAP is to be combined with the adult-oriented OMCS, incorporating the in-game validation, and possibly aimed at both adults and children, a varied and broad collection of useful commonsense knowledge could be gathered. To what degree this combination would be more varied and possibly broader than each of the two individual resources is a matter of further research.

## References


Ball, G.; Ling, D.; Kurlander, D.; Miller, J.; Pugh, D.; Skelly, T.; Stankosky, A.; Thiel, D.; Van Dantzich, M.; and Wax, T. 1997. Lifelike computer characters: The persona project at Microsoft research. In Bradshaw, J., ed., *Software Agents*. Menlo Park, CA: AAAI/MIT Press. 191–222.

Inhelder, B., and Piaget, J. 1958. *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.

Kuipers, B. 2004. Making sense of common sense knowledge. *Ubiquity* 45(4):13–19.

Liu, H., and Singh, P. 2004. Conceptnet: A practical common-sense reasoning toolkit. *BT Technology Journal* 22(4):211–226.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Michie, D., ed., *Machine Intelligence*, volume 4. American Elsevier.

McCarthy, J. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*. London, UK: Her Majesty's Stationery Office.

Minsky, M. L., and Papert, S. A. 1988. *Perceptrons: Expanded edition*. Cambridge, MA: The MIT Press. First published in 1969.

Minsky, M. L. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, NY: Simon & Schuster.

Schrooten, W., and Vermeer, A. 1994. *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen*. TUP(Studies in meertaligheid 6).

Singh, P. 2002. The open mind common sense project. Technical report, MIT.

Singh, P. 2003. Examining the society of mind. *Computing and Informatics* 22(5):521–543.

Speer, R.; Krishnamurti, J.; Havasi, C.; Smith, D.; Lieberman, H.; and Arnold, K. 2009. An interface for targeted collection of common sense knowledge using a mixture model. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, 137–146. New Brunswick, NJ, USA: ACM.

Stork, D. 1999. The Open Mind initiative. *IEEE Expert Systems and Their Applications* 14:16–20.

Van Mulken, S.; André, E.; and Muller, J. 1998. The persona effect: How substantial is it? In Johnson, H.; Nigay, L.; and Roast, C., eds., *People and Computers XIII: Proceedings of HCI-98*, 53–66.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systemsn Human Factors in Computing Systems*, 319–326. New York, NY: ACM.

Von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51:58–67.

Von Ahn, L.; Kedia, M.; and Blum, M. 2006. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 75–78. New York, NY: ACM Press.

Vossen, P.; Hofmann, K.; de Rijke, M.; Tjong Kim Sang, E.; and Deschacht, K. 2007. The Cornetto database: Architecture and user-scenarios. In Moens, M.-F.; Tuytelaars, T.; and de Vries, A., eds., *Proceedings of the Seventh Dutch-Belgian Information Retrieval Workshop (DIR 2007)*, 89–96.