A Kid's Open Mind Common Sense

Nienke Eckhardt
HAIT Master Thesis series nr. 08-07

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Communication and Information Sciences,
Master Track Human Aspects of Information Technology,
at the faculty of humanities
of Tilburg University

Thesis committee:

Prof. Dr. A.P.J van den Bosch
Prof. Dr. E.O. Postma

Tilburg University
Faculty of Humanities
Department of Communication and Information Sciences
Tilburg, The Netherlands
December 2008

Abstract

This thesis explores whether children can be a reliable resource for adding commonsense knowledge to the Open Mind Common Sense database. To gather commonsense information from children, we have created a game based on an existing commonsense game called *Verbosity*, in which participants are asked to describe and guess different nouns. We have derived several nouns from a corpus drawn from reading material used at all classes in primary schools. The nouns that are presented to the children are split up into concrete and abstract words and into three levels of difficulty. Since the goal is to add the gathered data to the already existing Open Mind Commons Dutch website, participating children are asked to chose among six relationship templates (such as "X is a kind of Y") taken from the OMCS Dutch site. The reliability of the descriptions is checked by looking at the amount of correctly given descriptions, the vocabulary level and language use of the children.

In a period of three weeks time, a total of 123 children played our game. Dozens of children have played the game more than once, so playing a simple game seems to be much fun for children. The vocabulary knowledge of children aged ten to twelve is sufficient enough to be able to describe our assigned words. Next to this, most of our participants are in Piaget's Concrete Operational stage. This means that they are more engaged in here-and-now situations and more with concrete facts than adults are, this can be beneficial for describing words. 10-year-olds have indeed described more concrete and less abstract nouns than the older children. The children's descriptions were reliable and the amount of spelling errors and typographical errors was low. The difficulty level and concreteness of the words hardly made any difference in correctness of the description, so children can play this game with any kind of noun derived from the corpus, and will describe the word correctly most of the times. If this game were to be actually used to built a OMCS-Dutch corpus, much reliable knowledge could be added in a relatively short amount of time.

Preface

This thesis is the final project of my studies in Human Aspects of Information Technology at Tilburg University. In the beginning of 2008 I was invited to join the Dutch Common Sense team for the Battle of the Universities. The plan was to create a Dutch version of the Open Mind Common Sense website. Unfortunately, our team did not make it to the next round. However, development of OMCS-Dutch continued and together with BDM bachelor student Pim Nauts, we were given the chance to set up a new and original experiment about adding commonsense knowledge from children to the OMCS-Dutch database.

I want to express my gratitude to my friends, family and fellow students, whose support kept me going through the writing of this thesis. I would particulary like to thank my supervisor, professor Antal van den Bosch for his guidance and advice. His knowledge and insights in the field of Artificial Intelligence and computational linguistics seem to be endless. I would also like to thank the following people who have helped us during the setup of our experiment: thank you Rintse van der Werf and Peter Berck for giving us advice about children's cognitive development and helping us with the technical aspects of the game, children's day care center "SSK-Spring" in Boxmeer for giving us the opportunity to perform our pilot test there and Marga van Zundert from the Kinderuniversiteit for helping us getting hundreds of enthusiastic participants. Further gratitude goes out to professor Eric Postma for wanting to participate in the exam committee. Special thanks go out to my boyfriend Franklin, who not only supported me through stressful times, but also helped me out with the data preparation when I got lost in all of it. And final thanks go out to my mother, who is always there for support, even through the most difficult of times.

Contents

Chapter 1

Introduction

A window is usually transparent, walls are not. Liquids can be poured into a glass, solids cannot. If you cover an object with your hands, you know it will still be there even though you can not see the object. This is basic commonsense knowledge. Common sense is regarded as the kinds of facts and concepts people would commonly agree on and hence do not talk about generally in everyday communication. Most of the commonsense facts we know are learned unconsciously or learned from experience. This is why it is so hard for adults to explicit commonsense information; it is stored there to help us function properly in our daily lives but it is not stored there to be uttered. However, all of us have been at an age when we did not possess all this commonsense knowledge.

Children seem to make a sport out of asking ”Why..?” about so many facts of life, that it sometimes even tires out their parents. They seem to have an innate curiosity for all these facts of life which appear so seemingly obvious for adults. Jean Piaget, a pioneer in children's cognitive development, supports this innate curiosity by stating that children actually think differently from adults. They are *”active builders of knowledge-little scientists who are constantly creating and testing their own theories of the world”* [Papert, 1999]. And while they are building their knowledge of the world, they are building their commonsense knowledge at the same time.

Unlike children, computers are unable to constantly test their (pre-programmed) theories. Computers have no idea how to reason with commonsense knowledge. They only know how deal with problems that lie within the boundaries and goals that the programmer has developed; when they are presented with new problems, they are unable to use simple logic to achieve their goals. Instead, computers excel in solving many complex issues such as beating chess players or assisting surgeons during open heart surgeries. Computers are great at tasks like these because they simply perform the task they are told to do. However, they are unable to tell you to take your umbrella with you when you go outside because it is probably going to rain. In short, there is a difference between domain-specific knowledge and general knowledge. Although Artificial Intelligence (AI) research has been quite successful at creating systems with domain-specific knowledge (as the aforementioned chess programs), a truly *intelligent* machine that can reason about all aspects of everyday life has never been created. Putting everyday commonsense knowledge into computers has been a long-standing dream in the world of AI because this is what could create true intelligent machines. To enable this, all the basic commonsense facts of life should be collected and put into computers. Estimates say the average person knows tens to hundreds of millions of facts. This would mean that all these facts need to be collected [Minsky, 2006]. Researchers have now found inspiration in community-based websites such as Wikipedia[1] and Delicious[2], in which the general public generates the contents of the website by adding information. In Wikipedia, users are motivated to add and correct information with the idea in mind that they are contributing to a large database of interesting information. Delicious is a *social bookmarking site*, in which users share their internet bookmarks with each other. By adding tags and descriptions to bookmarks, users add extra information which can be used to look up certain sites. Delicious can recommend sites to users based on the users' bookmark collection. Hundreds of thousands of pieces of information have already been added to these sites by its users. AI researchers now state that if millions of persons would express their commonsense knowledge on a website, then there is an enormous amount of knowledge that can be gathered to make computers "smarter". Millions of commonsense facts have already been gathered through websites and other commonsense projects. However, many hundreds of millions of facts are needed. This is why besides collecting facts through websites, commonsense-collecting games have been developed to motivate the public to explicit common sense. However, these games were mainly targeted at adults and because of the simple gameplay adults can become bored quickly.

This research is based on the idea that since children have a natural curiosity and interest in these facts, it could be a good idea to see children as a new target group for collecting commonsense information. Also because of that interest, children could be more motivated than adults in trying to explicit commonsense information. Besides, children enjoy playing online games, so playing an online commonsense game with children could be a success as well. Previous research usually focused on either commonsense reasoning for computers (with or without the help of the general public) or cognitive and commonsense development in children. Our research aims at combining these two subjects, which could be a potential wealth of commonsense information for these databases.

1.1 Research Question

Children's natural curiosity in facts can provide to be a very productive resource for adding information to a commonsense database. The commonsense database we are using for this research is that of the Open Mind Common Sense (OMCS) project, on which we will elaborate more in Section 2.3.4. We want to find out whether children can add valuable and reliable commonsense information to this database. To create a large enough data set, we need enough (enthusiastic) subjects who are willing to enter data. The

[1] http://www.wikipedia.org
[2] http://delicious.com/

original interface from the OMCS website is not very appealing to children; it is quite dull and corporate looking and besides that, the interface would be too complicated to handle. This is why we have created an online game, based on an existing commonsense game called *Verbosity* [von Ahn et al., 2006]. Verbosity makes use of the knowledge that everyone possesses by asking internet users to describe several concepts in a game. The reliability of the descriptions are checked by showing other users these descriptions while they have to guess the concept that was described in these descriptions. If the concept is correctly guessed, the descriptions are marked as reliable. Our game will also ask the children to describe several words and to guess a word by showing the children descriptions from other children. The difference between Verbosity and our game is that this game can only be played by one person, without that person being dependent on another player. By doing so we want to answer our main research question:

Can children (aged 10-12) be a reliable source to add information to a commonsense database?

To find out whether children are indeed a reliable source for commonsense data, the subjects' answers will be analyzed on several different elements. We will first provide an overview of commonsense and cognitive development in children and their level of vocabulary knowledge at age ten to twelve to find out whether children of that age are indeed able to produce correct descriptions of words. Then we want to know whether they enjoy playing the game. This will be asked by children in a pilot test and whether or not the game is played more than once. Also, the correctness of the descriptions will be checked. The children will be shown a selection of nouns, which have been split up into difficulty level and whether they are concrete or abstract. Concrete nouns are nouns which are physical objects that can be sensed, while abstract words are concepts which you cannot observe. More information on concreteness can be found in Section 3.1.4. Next to that, we have used six existing relationship templates from the OMCS site, in which the children can put their answers. The preferences for these relationships and the kind of descriptions that were put into them are also checked. We will also research details of the given descriptions: the kind of assigned words the children prefer to describe and next to that, the kind of words children use in their descriptions.

Not much research has been done on this topic. This thesis sets an explorative step. Its aim is to answer the main research question by analyzing and mapping out a range of findings and issues that are associated with making use of children to collect commonsense information.

1.2 Outline

In chapter 2 we will explain two different aspects on Common Sense: the first part of this chapter deals with common sense development and cognitive development of children. The second part of chapter 2 gives an overview of giving computers common sense. Chapter

3 deals with the preparation of the experimental data and the choices that were made concerning the development of our commonsense game. In chapter 4 the results of the experiments are discussed. The conclusions of our research and points for further research are discussed in chapter 5.

Chapter 2

Common Sense

In this chapter, we describe commonsense and cognitive development of children, and we examine past and current states of affairs in AI research at creating systems with common sense. Even though these two subjects seem very different from each other, they both form the basis for our experiment.

2.1 Commonsense Development in Children

Some basic commonsense knowledge is illustrated in the following example, taken from The Emotion Machine by Marvin Minsky:

"A package that is tied up with string.":

> *With a string you can pull, but not push, a thing.*
> *If you pull too hard, a string will break.*
> *You must fill a package before tying it up.*
> *Loose strings tend to get tangled and knotted.*
> *[Minsky, 2006, p. 163]*

Usually, people never look as consciously at situations as described above. This information is always taken for granted because it is basic common sense. Many abilities and knowledge adults now take for granted are acquired during their earlier childhood. We can easily remember many events in our life, but why are we unable to recall events from our earliest childhood? Minsky presumes that this *Infantile Amnesia* [Minsky, 2006] is caused by the fact that humans in their early childhood have not yet developed the skills to actually remember these memories. Yet, this earliest childhood (from newborn to three years old) is the period in which some of our commonsense knowledge and most of our basic abilities are acquired. When we reach the age at which we are able to consciously recall past events of our lives, memorization and recalling memories have become automatized skills; we are unable to explain how we memorize.

2.2 Cognitive Development in Children

Since we can never be absolutely certain of what exactly happens in the minds of children, we can only see the outcome of these internal processes. Several different views on children's cognitive development have come and gone in the last decades. In his book "*The Society of Mind*"[Minsky, 1988], Minsky explains his view on mental development; a child learns from experiences, but also has some innate abilities to give basic structure to these experiences which, in time, will help the child develop different kinds of reasoning. To become more intelligent, a mind cannot simply accumulate more and more knowledge; it has to improve the ways of how already established knowledge is managed. Minsky calls this Papert's Principle:

"Some of the most crucial steps in mental growth are based not simply on acquiring new skills, but on acquiring new administrative ways to use what one already knows."

Seymour Papert has done much research on the cognitive development of children with help of computers. He is a great advocate of educational computer-based technology. With this in mind, Papert created a programming language named "Logo" to help children develop their problem solving skills. And indeed, many of the children that learned to program with Logo, increased their logic skills and scored higher in maths [Papert, 1980]. Papert's view on children's cognitive development is largely based on the experiments of child psychologist Jean Piaget, whom he worked with. As said in our introduction, Jean Piaget was a pioneer in the field of cognitive development in children. In the early and middle part of the 20th century, Piaget conducted many experiments with children and his results provided a framework for understanding children's cognitive development. His theory tries to illustrate what children are capable of at different ages. They start out with a completely different view of the world from the view that adults have. He found out that children at a certain age constantly gave wrong answers to questions that older children answered correctly. In fact, even when they had already heard the correct answer, younger children still continued to give wrong answers. Figure 2.1 illustrates that when, in front of the children, the liquid from the short, wide jars was poured into the tall, thin jar, a typical 5-year-old child will say that the tall, thin jar will always contain more water, whereas a 7-year-old understands that it is the same amount. From this "illogical stubbornness" he concluded that the cognitive processes of young children are inherently different from those of adults, and that children's views of the world are constantly evolving, even far into adolescence [Ackermann, 2001]. His theory contrasted with two dominant views of cognitive development; one stating that babies minds are like "blank slates" which would passively learn everything from the environment, and the other stating that babies are born with innate knowledge (nurture versus nature). Piaget stated that knowledge is not simply delivered to a child, but is acquired through interaction with the world. His ideas and findings laid out the foundations for the theory of constructivism.

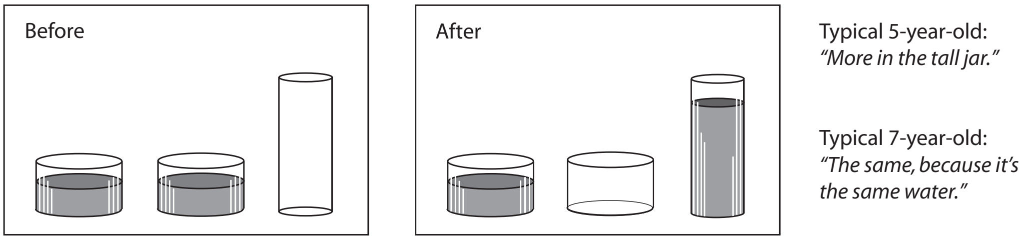


Figure 2.1: Example from one of Piaget's experiments which showed that children think differently from adults.

Piaget suggests that there are two major principles that stimulate intellectual growth: *organization and adaptation*. Organization refers to the innate capacities of babies to structure the information that they receive. Piaget called the basic building blocks of giving structure to information *schemata*. These are sets of interrelated features which an individual associates with a concept. An example of a schema of the concept *Table* is shown in Figure 2.2.

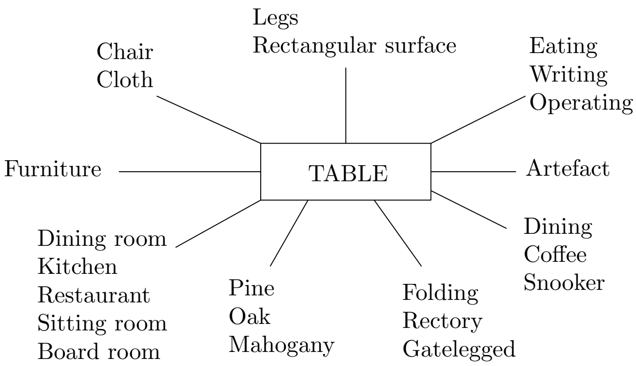


Figure 2.2: A schema for the concept *Table*. Source: "Psycholinguistics" [Field, 2003, p. 39].

Every individual makes use of schemata to understand the world, and this starts from birth. Piaget's second principle that stimulates intellectual growth is called adaptation. Piaget split this principle up into two different aspects. The first aspect of adaptation is called *assimilation*. This means that children need to adapt to stimuli from outside to be able to survive in an environment. For example, when a child is constructing a schema of the concept "table", the first aspects it will construct in the schema will probably be: "kitchen, table, dining". Later, when the child sees that a table can also be used for writing and drawing, it has to accept this new facet and this will then be added to the already existing schema through assimilation. The new information can be fitted to already existing categories in the child's mind. The second aspect of adaptation is *accommodation*. For example, when a child sees a pool table for the first time, the child will understand that a pool table has many of the same aspects of a dining table, but it also has some large discrepancies with the child's own schema of the concept "table" ("kitchen, table, dining, drawing"). The child will have to partially reconstruct its schema about "table":

it has to accomodate its original ideas to fit to the new realities of the concept "table" [Bhattacharya, 2008]. These two processes of organization and adaptation internalize the children's experiences with the outside world.

According to Piaget, children go through four subsequent stages in their cognitive development. Children start out thinking very differently from adults, but as they grow up, they start to adopt the same mental routines as adults have [Molnar, 2005] [Atherton, 2008].

1. **The sensori-motor stage.** *(Birth - two years old)* - The child will learn to understand the world through trial and error (by using its senses and motor skills).

2. **The pre-operational stage.** *(Two - seven years old)* - Use of language, memory and imagination, cannot see things from other points of view (egocentric).

3. **The concrete-operational stage.** *(Seven - eleven years old)* - The child can think logically, but only related to concrete objects and events, child becomes less egocentric.

4. **The formal-operational stage.** *(Eleven years old and up)* - Use of symbols for thinking about abstract concepts and relationships, starts to think about the future.

Piaget believed that all children pass through these phases in chronological order. Some processes cannot be learned until other processes become available. The ages at which these stages take place can differ per child [Bhattacharya, 2008]. Minsky illustrated this mental growth by stating that "you cannot start to build a house by placing its roof on top; first you have to build some walls" [Minsky, 1988, p. 179]. We can illustrate this by the fun babies have when parents play *peekaboo* with them; the parent hides his/her face behind his/her hands and suddenly removes them and says "peekaboo!". All infants enjoy this magical reappearance of the parent, because they still have no notion of *object permanence*: being aware that objects continue to exist even when they are no longer visible[1]. This awareness takes place after around six months of age. This example illustrates that it is now basic "common sense" for adults to expect that a persons face is still there when that person puts his/her hands in front of it, while there once was a time where we did not possess this piece of common sense. Children use these acquired "basic" principles to assimilate new information into their internal mental states. Like building the foundation of a house before even beginning to build the walls and roof; every newly adapted piece of information will be the foundation for creating more complex mental capabilities. What this means is that what adults see as basic common sense, are actually facts that we all had to learn sometime in our early childhood. And because of infantile amnesia, it seems to adults that these commonsense facts have always been present in our minds; we simply cannot remember the moment when this knowledge was acquired. Other commonsense facts, such as picking up a phone when it rings or knowing that paper is made from wood,

[1] `http://www.britannica.com/EBchecked/topic/423787/object-permanence` Retrieved: November 18th, 2008

are acquired at a later age and which are facts from which we sometimes do still remember how we learned them.

The cognitive development of the subjects in our experiment will be somewhere in between the *concrete-operational* stage and the *formal-operational* stage, because their ages range from ten to twelve years old. This means that the children will not explicit "magical" descriptions of concepts or that they will only describe concepts from their own point of view because they have already outgrown that phase. In terms of commonsense development of our subjects it can be stated that at their age, these children have already acquired many of the basic commonsense rules. Even more, children in the concrete-operational stage think less abstract, which may imply that they will still experience fun when they have to describe basic, concrete words in the game we designed for our study.

2.3 Giving Computers Common Sense

Computer scientists have been trying to create computers with "Artificial Intelligence" (AI) for at least 60 years now. In 1950, Alan Turing described his now infamous "Turing Test" in which a human judge had to chat with a human and a computer. This judge had to distinguish the computer from the human. If the judge could not find any difference between the two, the machine had passed the test and was considered "intelligent". One can imagine that there has been disagreement as to what "passing the test" means, because imitating something or someone does not mean that a machine is indeed "intelligent". Almost 60 years later, there are computer programs that can assemble cars in factories, diagnose heart attacks and solve complicated logistic problems without too much effort. Yet we have never seen a computer that can properly recognize a photo or lay a brick wall. Even nowadays there is an annual contest based on the Turing Test: the "Loebner Prize"[2]. This contest awards the most human-like computer with a bronze medal. The gold medal (and a money prize of $100,000) is only reserved for the computer which has indistinguishable responses from those of a human. To date, no-one has ever been awarded that gold medal. Several renowned scientists in the field of AI like Marvin Minsky [Minsky, 2006] and John McCarthy [McCarthy, 2007] blame this lack of intelligence of computers on the fact that computers lack common sense. Minsky adds that programmers only tell programs what things to do, without telling the program why we actually want them done. He also states that a typical program will quit when its method fails or when it lacks knowledge, while a person is resourceful enough to find an alternative.

If we apply "Papert's Principle" (see Section 2.2) to the field of AI, machines should first be equipped with ways to process information correctly. Next, it should make inferences of all the data it "experiences", and it should then improve its ways of processing that information by constantly checking if it is in accord with previous experiences again and again until you can say it is truly 'intelligent'. This is the basic idea behind many recent AI projects.

[2] http://www.loebner.net/Prizef/loebner-prize.html Retrieved: December 3rd, 2008

2.3.1 How Much does the Average Person Know?

Many researchers now believe that to create a truly intelligent system, computers first need to learn many little facts about all common aspects of life. As said in Chapter 1, many of the commonsense facts we know are so obvious that we take them for granted. But how much information do we actually store in our minds? Several estimates have been made about the amount of common sense that people possess. To measure the size of our knowledge, Thomas Landauer makes use of the smallest piece of information we know; a "bit". He states that human commonsense knowledge is probably around 1,000,000,000 (one billion) bits in size [Landauer, 1986]. Minsky suggests that if we take our language system, object knowledge and social realm into account, a "humanlike reasoning system" requires hundred millions of items of knowledge [Minsky, 2006]. However, in an interview in the International Herald Tribute (as cited in [Liu and Singh, 2004]) Minsky mentions that "*Common sense is knowing maybe 30 to 60 million things about the world.*". Other estimates also come up with numbers in the order of hundreds of millions of commonsense facts [Singh, 2002]. Even though these estimates do not specify how complex these "basic common sense" facts are, and of how many basal elements these facts consist, we can easily state that there are very many facts to gather.

2.3.2 Creating a Commonsense Database

The idea behind all commonsense databases is that they need to contain fundamental assertions such as "*You have to be awake to eat.*" Statements like these are unlikely to be published in books, websites and encyclopedias because they are so obvious to us human beings. However, these statements need to be put into words before a computer can learn from it. Several successful and less successful attempts at giving computers commonsense knowledge have been made in the past years. Many of the projects that failed did so because most of the time people misjudged the amount of work that has to be done to enter all these millions of facts into a database. However, more recent projects such as OMCS have gathered a lot of useful knowledge which already is being used in (experimental) applications.

2.3.3 Research on Commonsense Databases

Cyc

One of the largest and well known projects that is based on building a collection of commonsense knowledge is the Cyc Project, started in 1984 by Douglas Lenat [Lenat, 1995]. So-called "knowledge engineers" manually insert knowledge into the database via a specific formal language called CycL. Cyc will automatically connect the inserted assertions and create inferences, which can be a start at getting computers to reason about statements. Researchers of the Cyc project have collected and inserted over a million of these commonsense facts over the past years which makes the data and the inferences that can be made with Cyc very reliable. However, since we seem to know tens to hundreds of million of facts about the world, not even ten percent of our commonsense knowledge has been put into the Cyc database. Even though Cyc is a very ambitious project and has gathered much useful knowledge, there are also a few limitations. Cyc's knowledge is presented in a logical framework which is great for deductive reasoning, but this formalized, hierarchical framework makes interpreting natural language and use of different contexts impossible. Natural language is full of ambiguities, which have not been taken into account enough by Cyc [Roush, 2006]. In contrast, ConceptNet, a commonsense knowledgebase based on data gathered by the OMCS website, excels at contextual commonsense reasoning over natural language because its foundation is based on sentences created by the general public [Liu and Singh, 2004]. Another disadvantage is the fact that you need experts to add knowledge to the database. This is an almost impossible task when hundred millions of facts need to be added. Next to this, Cyc has been a very expensive project, with a spending budget that is running in the tens of millions of dollars.

ThoughtTreasure

In 1994, Erik Mueller started a commonsense project called "ThoughtTreasure". This project is partly based on Cyc's knowledge representation and limited use of natural language. As with Cyc, commonsense knowledge has to be formulated and represented in a specific formal language before adding it to the database. ThoughtTreasure contains about 100,000 pieces of common sense existing of declarative and procedural knowledge, which is considerably less than Cyc [Mueller, 2003]. Commercial use of ThoughtTreasure did not turn out to be a success and new developments have not been made since 2000, apart from some data that has been added to the database.

MindPixel and Open Mind Common Sense

Mindpixel is an internetbased project developed by Chris McKinstry in 2000. Knowledge could be added by every visitor of the site. Participants would use normal sentences to explicit commonsense knowledge. This use of natural language when describing concepts lead to a large database with 1.4 million so-called "mindpixels" in 2004. Compared to Cyc, this is a much faster grow of facts added to the database. In 2005, Mindpixel lost its free server and the project was put to a hold by McKinstry, and after his untimely death in 2006 Mindpixel's development was brought to a halt. Nowadays, the site has a game where (mostly irrelevant) statements taken from the Mindpixel database are asked to be true or false.[3]

Another project which has been developed around the same time as Mindpixel and which is also based on using internet users for adding data is the "Open Mind Common Sense"

[3]Source: Wikipedia *http://en.wikipedia.org/wiki/Chris_McKinstry* Retrieved: November 3rd, 2008

project (OMCS) [Singh et al., 2002]. The idea behind OMCS is the same of that of Mindpixel: every layperson can add commonsense knowledge to the database because it is knowledge that even children possess [Liu and Singh, 2004].

Mindpixel and OMCS were both pioneers in collecting commonsense knowledge by asking the general public to explicit their own commonsense knowledge. Both projects have a collaborative system in which users of the site create and classify statements as true or false, and by this they build up a large collection of commonsense facts. The reliability of the statements is monitored by other users who can rate statements as true or false.

Other projects such as Cyc and ThoughtTreasure were dependent on experts to add commonsense data to their database, which is a slow and expensive process. By using the general (internet) public, a wealth of commonsense knowledge is available at the tip of the researchers' hands and for a fraction of the costs when using experts. In our opinion OMCS is the most promising among the mentioned projects. Since our research is largely based on OMCS, we will talk more extensively about OMCS in the next section.

Verbosity

One of the latest developments for collecting commonsense data is gathering "free labor" by letting people play games to collect the data. Luis von Ahn developed an image-guessing game called *the ESP game* in which two anonymous individuals were paired up and were shown an image. Both persons had to write down words to describe the image. If both persons would enter the same word, they would get points [von Ahn and Dabbish, 2004]. Inspired by the ESP-game, von Ahn developed *Verbosity* in 2006 [von Ahn et al., 2006], a game for collecting commonsense facts. Verbosity's aim is to appeal to a large audience, regardless of whether or not they are interested in contributing to AI. Instead of asking users to enter a statement or rate such statements, like in OMCS and Mindpixel, the developers of Verbosity wanted users to play a game inspired by the popular party game *Taboo*.

The gameplay of Verbosity is as follows: two players are randomly selected. One of the players is chosen as "Narrator" and the other is the "Guesser". The Narrator is shown a secret word and he/she needs to get the Guesser to type that word. The narrator does so by sending hints (in the form of sentence templates) to the Guesser. For example, the Narrator sees the secret word *Milk*. He or she can then choose between several templates to give a hint, a hint could be "..is typically near *cornflakes*". If the Guesser would then enter *Milk*, both players would get points and the system would assume that the given hints from the Narrator are accurate, and stores them into the database. By taking the time it takes to correctly guess a word and the random pairing of the players into account, the accuracy of the hints is ensured even more.

Next to the two-player game, there also is the possibility to use only one player. By emulating the Narrator and by displaying a set of previously collected facts, the (human) Guesser could then still guess a word, and the reliability of the facts can be verified.

Emulating the Guesser is more difficult, because to protect the illusion of a real game taking place, you cannot let the computer guess the concept at once, or let the computer utter totally unrelated concepts first before guessing correctly. The developers of Verbosity have tried to intercept this problem by compiling a list of related words which help the computer to guess a word more like a human would.

WordNet

WordNet is an electronic lexical database which has organized English nouns, verbs, adjectives and adverbs into synonym sets of which each set represents a distinct meaning of a word. WordNet's development was inspired by psycholinguistic theories of human lexical memory and was developed to offer a more intuitive dictionary and thesaurus which could also support automatic text analysis. Development of the lexicon began in 1985, initiated by professor George A. Miller of Princeton University. The project is currently overseen by Dr. Christiane Fellbaum [Fellbaum, 1998]. Currently, the database contains over 155,000 words, over 117,000 synonym sets for a total of almost 207,000 word-sense pairs[4] that have been added to the database by knowledge engineers. Like Cyc, WordNet also structures its knowledge hierarchically by using formal taxonomies of words: it uses hypernyms/hyponyms and IsA relationships to name concepts. For example, a hypernym for "table" is "furniture".

In the Computational Linguistics community it is one of the most popular and widely used semantic resources. WordNet is so successful because of its simple semantic structure with words at the nodes. With little effort, WordNet's data can be used for automatic text analysis [Liu and Singh, 2004]. WordNet is a bit of an outsider compared to the abovementioned knowledge databases, because its background does not stem from AI but from the field of psycholinguistics. However, WordNet is often used for AI purposes, for example in combination with existing commonsense databases.

2.3.4 Open Mind Common Sense

In 1999, David Stork started the "Open Mind Initiative" (OMI). This initiative is based on acquiring commonsense knowledge through collaboration. The goal of this project is to find out what issues arise when using the general public for knowledge acquisition instead of expert knowledge. With this in mind, OMI wants to create intelligent software. By using non-expert knowledge from the general public, there is an enormous source of knowledge directly available to be implemented into a commonsense database [Stork, 1999]. The knowledge that OMI gathers can serve as a basis for intelligent software. Many side projects have already originated from this initiative, such as "Open Speech" (a speech recognition tool), "Open Mind Handwriting Recognition", and "ARIA" (software that manages photos with use of common sense). OMI's largest project is the aforementioned Open Mind Common Sense project. This project, initiated by Marvin Minsky and Push

[4] http://wordnet.princeton.edu/man/wnstats.7WN, Retrieved: December 11th, 2008

Singh, has been collecting data and is the basis for many different commonsense projects ever since its start in 1999. OMCS has already collected a large amount of data (around 700,000 English facts from 14,000 contributors in 2004) in a relatively short time and for a fraction of the cost of projects like Cyc. This data has now been used to build a large semantic network called *ConceptNet* [Havasi et al., 2007]. One of the points of criticism of OMCS is that natural language contains too much ambiguity to be actually used to reason with. However, the OMCS developers claim that these ambiguities can be solved by looking at the context of the contributions [Liu and Singh, 2004].

Collecting Commonsense Facts

The goal of OMCS is to create a knowledge collaboration network which can be easily understood and used by the general public. All commonsense data in OMCS is gathered by users of the web who voluntarily share their commonsense knowledge by describing concepts. OMCS has been under constant development since 1999. In the first version of OMCS, OMCS-1, users could add knowledge in three different ways [Singh, 2002]:

- By stating **facts**. (Ex: *People want to be warm. Cats are mammals.*)
- By filling in **templates**. (Ex: *A hammer is for _____. Somewhere you find a bed is _____.*)
- By typing in knowledge in a **free-form** way after reading a short story. (Ex: *Bob had a cold. Bob went to the doctor → Bob was feeling sick. Bob wanted to feel better. People with colds sneeze. The doctor made Bob feel better. Etcetera.* (The "→" indicate deductions people have made from the short story.)
- By adding **descriptions** of photos or other visual events. (Ex: *The small red ball rolls past the big blue ball.*)

In a 2002 study human judges evaluated a sample of the OMCS corpus. First evaluations for measuring *truth*, *sense* and *neutrality* of the data entered were reasonably positive; over 75% was rated in favor of these three attributes, which means that most of the data is truthful, makes sense and is not overly biased. Since the added data has not been put into a certain form of logic, the OMCS team had to use different extraction techniques to extract commonsense knowledge from this data. Part-of-speech taggers were used to automatically assign the parts of speech such as verbs and nouns to words. Syntactic parsers were used for automatically assigning the syntactic structure (the rules that form words into sentences) to the sentences that people have entered. From this data, the team developed patterns in which the raw data could be formed into basic representations to be used in later versions of OMCS [Singh et al., 2002].

Knowledge Entry in OMCS

With the free-form way of adding knowledge in the first version of OMCS, the team could create templates for the improved version of OMCS: "Open Mind Commons". By asking contributors to add knowledge in this structured manner, the knowledge extraction procedure would be much more simpler. However, free-form knowledge is still allowed in Open Mind Commons, because the team did not want to miss out on any other commonsense knowledge. In Table 2.1 a selection of the template predicates used in OMCS are shown. The knowledge that is entered into the templates can represent either noun phrases, verb phrases, adjective phrases or prepositional phrases (when describing locations). Verb phrases can consist of complex structures with embedded prepositional phrases and noun phrases. In OMCS, the Verb phrases usually consist of a verb in combination with a noun, such as "make breakfast". Currently, OMCS contains around 20 relation types.

Relation	**Example sentence pattern**
IsA	*NP* is a kind of *NP*.
MadeOf	*NP* is made of *NP*.
UsedFor	*NP* is used for *VP*.
CapableOf	*NP* can *VP*.
DesireOf	*NP* wants to *VP*.
CreatedBy	You make *NP* by *VP*.
InstanceOf	An example of *NP* is *NP*.
PartOf	*NP* is part of *NP*.
PropertyOf	*NP* is *AP*.
EffectOf	The effect of *VP* is *NP*—*VP*.

Table 2.1: Selection of Predicate Templates used in OMCS.

ConceptNet

The relation types that are derived from the OMCS corpus are constantly added to the ConceptNet database. The semi-structured sentences taken from OMCS enable the researchers to extract the sentences into knowledge that is more computable. This means that the knowledge can be added to a semantic network, which is basically a large collection of concepts that have a specific meaning. These meanings are interconnected with each other, so that computers can reason with this knowledge. A semantic network such as ConceptNet consists of nodes which represent concepts (sometimes accompanied with a verb) and links which represent *predicates* (that part of a proposition that makes an assertion or a relation between objects). These predicates are extracted from the template relationships that OMCS users have filled in. This node/link structure is comparable with that of WordNet, apart from the fact that WordNet uses only words with atomic meaning and ConceptNet adds action verbs to its nodes. This allows ConceptNet to create knowledge of a greater range of concepts. However, because ConceptNet's words are not tagged like the synonym sets in WordNet, word senses are not distinguished. Figure 2.3 gives a network example of the concept "wake up in morning". This figure also shows that, unlike Cyc and WordNet, there is no hierarchy in the relations. Conceptnet only shows the semantic relations between concepts.

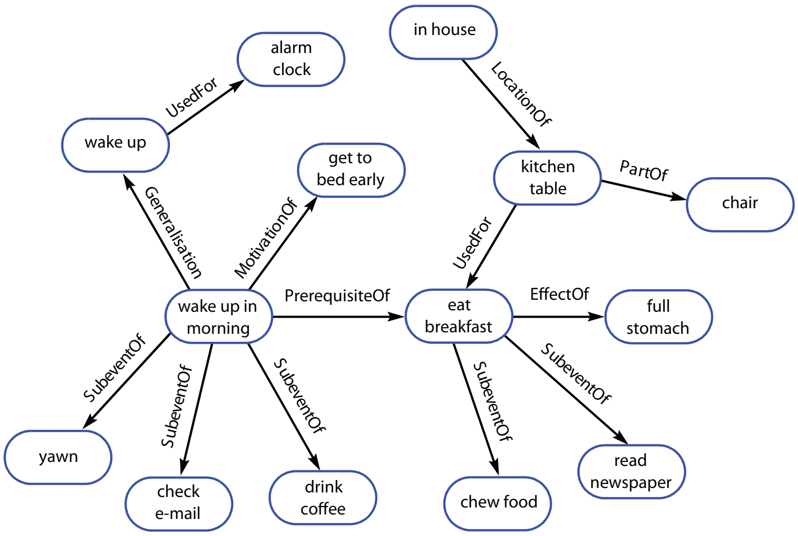


Figure 2.3: An example of ConceptNet's semantic network of commonsense knowledge. Source: [Liu and Singh, 2004].

Connecting concepts semantically makes informal reasoning over concepts possible, whereas this is impossible in a knowledgebase like Cyc. This is because ConceptNet takes the context of each concept into account. Where WordNet excels in lexical categorization and word-similarity determination and Cyc excels at careful deductive reasoning and for unambiguous situations, ConcepNet is superior at contextual commonsense reasoning over real-world text. This can help computers learn to make better use of context in sentences, and in the researchers' own words *"revolutionize textual information management"* [Liu and Singh, 2004, p. 4].

Open Mind Commons Dutch

In January 2008, the ILK research group at Tilburg University started collaborating with the developers of OMCS to create a Dutch version of OMCS: Open Mind Commons Dutch. The goal of this project is to create a large collection of commonsense knowledge in Dutch. The corpus that will be created by this can be of great use for projects that have until now only been possible for the English language. OMCS-Dutch functions in the same way as the English version. However, creating a Dutch version of ConceptNet will be a task for the future, because collecting knowledge has to start from scratch for this Dutch version. This is why (Dutch) children can help adding new knowledge quickly to the database.

Some screenshots of the Dutch OMCS website are shown in Figure 2.4, 2.5 and 2.6. Figure 2.4 shows an overview of descriptions that were made about the concept *mens* (human). These descriptions can be rated by either clicking a "thumbs up" or a "thumbs down" icon or when it is an inappropriate description, it can be flagged. Below at "Open Mind wants to know" users can state whether two combined concepts are comparable to each other or not.

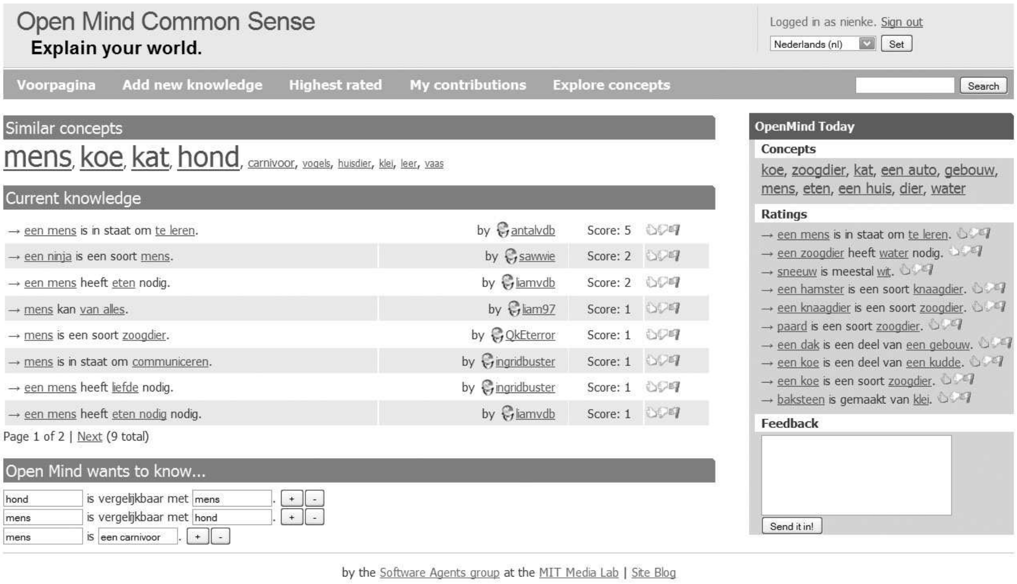


Figure 2.4: A screenshot of OMCS-Dutch which shows an overview of descriptions about the concept "mens".

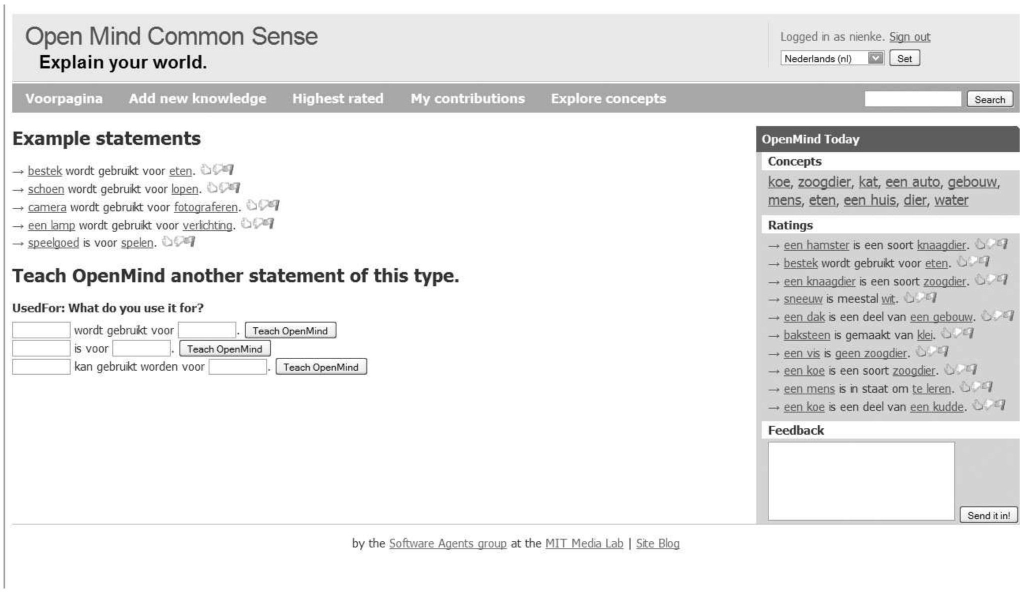
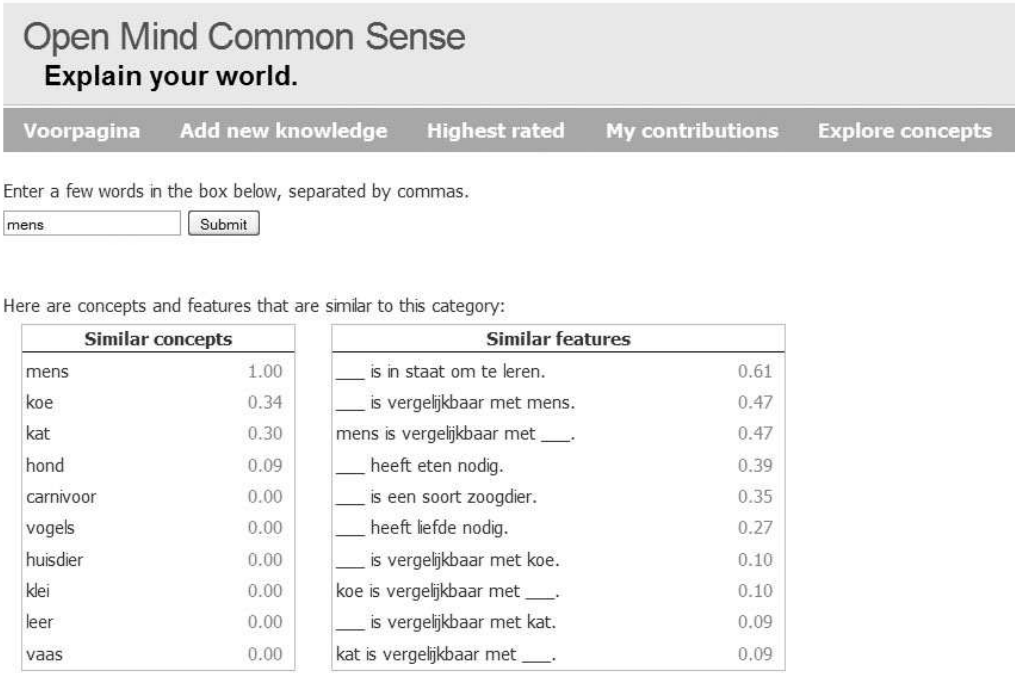


Figure 2.5: A screenshot of OMCS-Dutch which shows how to add new knowledge with the *UsedFor* relation.

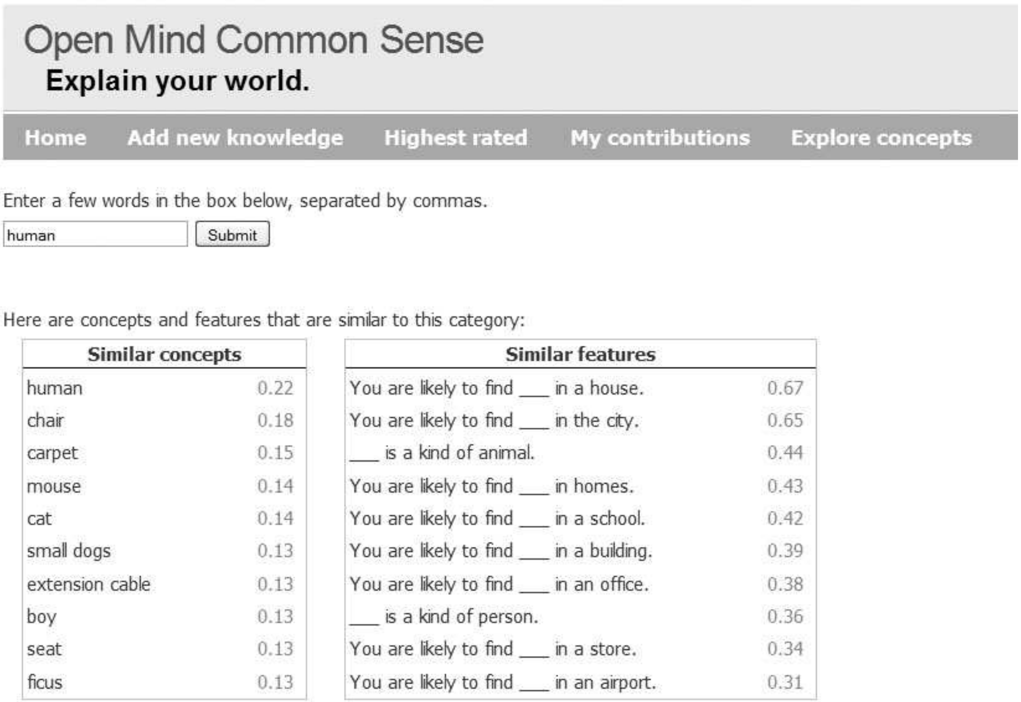
Figure 2.5 shows how the descriptions can be entered. This screenshot shows how users can describe a "UsedFor" relation. Figure 2.6 shows concepts which are in one way or the other seen as "similar" to the concept *mens* (human). Because not many statements have been added to OMCS-Dutch yet, strange concepts as "*koe is vergelijkbaar met* ___ " (cow can be compared with ___) score relatively high in the similar features list. When these results are compared with its English equivalent *human* in Figure 2.7, it is apparent that these similar features make more sense in English than in the Dutch version. In

OMCS-Dutch *mens* currently has received nine descriptions and in OMCS-English *human* has received 1,261 descriptions. Clearly, OMCS-Dutch still needs an enormous amount of knowledge to be added by volunteers.


Open Mind Common Sense
Explain your world.
Voorpagina Add new knowledge Highest rated My contributions Explore concepts
Enter a few words in the box below, separated by commas.
mens Submit
Here are concepts and features that are similar to this category:

Similar concepts		Similar features	
mens	1.00	___ is in staat om te leren.	0.61
koe	0.34	___ is vergelijkbaar met mens.	0.47
kat	0.30	mens is vergelijkbaar met ___.	0.47
hond	0.09	___ heeft eten nodig.	0.39
carnivoor	0.00	___ is een soort zoogdier.	0.35
vogels	0.00	___ heeft liefde nodig.	0.27
huisdier	0.00	___ is vergelijkbaar met koe.	0.10
klei	0.00	koe is vergelijkbaar met ___.	0.10
leer	0.00	___ is vergelijkbaar met kat.	0.09
vaas	0.00	kat is vergelijkbaar met ___.	0.09

Figure 2.6: Concepts and features of "mens" that are seen as similar by OMCS-Dutch.


Open Mind Common Sense
Explain your world.
Home Add new knowledge Highest rated My contributions Explore concepts
Enter a few words in the box below, separated by commas.
human Submit
Here are concepts and features that are similar to this category:

Similar concepts		Similar features	
human	0.22	You are likely to find ___ in a house.	0.67
chair	0.18	You are likely to find ___ in the city.	0.65
carpet	0.15	___ is a kind of animal.	0.44
mouse	0.14	You are likely to find ___ in homes.	0.43
cat	0.14	You are likely to find ___ in a school.	0.42
small dogs	0.13	You are likely to find ___ in a building.	0.39
extension cable	0.13	You are likely to find ___ in an office.	0.38
boy	0.13	___ is a kind of person.	0.36
seat	0.13	You are likely to find ___ in a store.	0.34
ficus	0.13	You are likely to find ___ in an airport.	0.31

Figure 2.7: Concepts and features of "human", English for "mens", that are seen as similar by OMCS-English

Chapter 3

Experiment Setup: Creating a Kid's Open Mind Common Sense

We have chosen to collect commonsense facts from children by letting them play a web-based game. Playing the game online will be at the expense of the reliability of the answers, because there will not be someone around to survey the subjects during the experiment. In our invitation letter, we have explained the participants (and their parents) that it is impossible to give incorrect answers and that mistakes are okay to make. However, it is to be expected that some children will ask their parents about a word or that they will "Google" the definition of a word. Some noise in the results will be expected because of these external factors. However, putting the game online gives us the opportunity to reach much more children and thus more useful answers than when conducting a supervised experiment.

3.1 Preparatory Data

To be able to let the subjects describe and guess different concepts, different nouns will be presented to them. These nouns will be split into difficulty level and whether they are concrete or abstract. As in Verbosity, we need to keep some structure in the descriptions of the subjects, but these descriptions should be usable for the OMCS database. This is why we have added relationship (predicate) templates that are also used in OMCS. If our game is to be used in practice for collecting commonsense facts, then the filled in relations can be directly implemented into the OMCS database. We will explain the selection of the nouns, their levels and the selection of the relationship templates below.

3.1.1 Word Selection

The words that are shown to the subjects are derived from a word list "Woorden in het basisonderwijs" (Words in primary school) that is developed by Schrooten and Vermeer

[Schrooten and Vermeer, 1994]. This list consist of over 26,000 words that are collected from different corpora (texts from novels, schoolbooks and recorded conversations) that are presented to children in primary education. These texts are divided into the grades in which they are presented. These 26,000 words are based on their lemmas (the "basic" form of a word), which means that the occurrence of "boats" will be put under its lemma, which is "boat". When we further use the definition "word" in this thesis, we mean the use of its lemma. The result of this is a reliable word list of the words which are presented to children during their entire primary school years.

The 200 most frequent words in the corpus are mainly function words (grammatical words which bear no particular semantics and which generally express a grammatical relationship with other (content) words; they "tie" our words together) such as "the, it, of, about" and auxiliary verbs such as "have, are". The 2,000 to 5,000 most frequent words below those are words that belong to our "basic" vocabulary knowledge which are commonly used in many different situations. The thousands of other remaining words are words that have relatively lower frequencies and tend to be more domain specific or more complex.

3.1.2 Noun Distinction

To describe different concepts, only nouns were picked from the word list. Compounds such as "schoenmaker" (shoemaker) or "woordenboek" (dictionary) were not selected, because these words consist of two (or more) lemmas combined, this can make the description of a word lengthy and which can produce circular descriptions such as "a shoemaker is a person who makes shoes". We have also tried to select a varied amount of words; many concrete nouns were species of animals and body parts, so we had to make a selection from these. The same goes for words that were more or less synonyms to each other. Proper names were also ignored.

3.1.3 Difficulty Level

The nouns that were first selected from the list were then split into different levels of difficulty. The difficulty level of a word is calculated in Schrooten and Vermeer's word list by measuring the *word frequency* (how often a word appears in a corpus) and by measuring the *spread* (whether the word appears very often in only one text or often in multiple texts). The more a word is presented to a child, the sooner the word is acquired. Less frequent words are therefore acquired at a later age. The value of this is represented by the *Geometric Mean* (further referred to as GM); the spread of a word is directly influenced by its score in the list. Words that are less frequent but do appear in multiple corpora receive a higher GM score than very frequent words that appear often in only one corpus. To determine whether a child should know a word or not, the GM should first be observed. In short we can say that the lower the GM of a word, the more difficult the word is. For our experiment, we have used the GM score of a word to split the assigned words into three levels of difficulty. Table 3.1 shows the geometric values and the amount of lemmas, in combination with the choice of grade in which these words can be offered to the children. For example, words with a geometric mean between 52-51 can be taught to children from (Dutch) grades three, four and five, because not all children are familiar with these words. However, it would be less effective to teach these children words with a GM higher than 52, because they already know their meaning.

Selected Lemmas (Cumulative)	**Geometric Mean**	**Choice of grade**
1000	> 115	1
2000	115 - 68	1, 2
3000	67 - 58	1, 2, 3
4000	57 - 53	1, 2, 3, 4
5000	52 - 51	3, 4, 5
6500	50 - 49	4, 5, 6
8100	49 - 48	5, 6, 7
12700	47	6, 7, 8
26590	46	7, 8

Table 3.1: Geometric values and amount of lemmas in combination with the choice of grade. Source: Guchte & Vermeer [Van de Guchte and Vermeer, 2003]

When we sort the complete word list on Geometric Mean, we can see if a child has seen the word often and should know it. For example, a word such as "paleis" (palace) has a GM of 66, and over 2,000 lemmas have a higher GM. This means that 9-year-old children, who have a vocabulary size of around 7,000 words (see table 3.4), would easily know the meaning of the word "paleis" because the child has seen the word regularly in his or her schoolbooks. The chance that a 9-year-old would know a word such as "dialoog" (dialogue) with a GM of 47 and 8,100 lemmas with a higher GM is much smaller, because this word is less frequently shown to the child.

Words with a geometric mean higher than 58 should all be familiar to children who are in the Dutch "groep drie" (first grade). We have labeled these words as *easy*; all our participants should understand these words very well. Words with a geometric mean between 48 and 58 should be words that children learn between "groep drie" and "groep zeven" (the first grade and fifth grade). These are labeled as *average*. Words with a geometric mean of 47 and 46 (which is the lowest geometric mean in the wordlist) are labeled as *difficult* which means that these words could be familiar to our participants, but there also is a possibility that our participants have never seen these words.

3.1.4 Concrete and Abstract

After the selected words were split into three difficulty levels, words were then manually classified as either *concrete* or *abstract*. Knowing the definition of a word is different from actually being able to describe a word. We all know how to describe the word *tree* in different ways, but imagine trying to explain a word such as *color*. This is why we have also distinguished *concrete* and *abstract* nouns. Concrete nouns refer to physical bodies which you can observe in one way or the other with one of your senses. Examples here are *animal, female* and *bicycle*. Abstract words refer to abstract concepts which cannot be sensed in any way, these are concepts such as *hate* or *love*. A noun defined as *easy* should be easy to understand for our subjects, but if it is an easy and abstract word, we expect these words to be much more difficult to describe.

Next to the manual selection, we have also automatically classified the selected words by using the Cornetto Database [Vossen et al., 2007]. The Cornetto Database is a combination of the Dutch part of EuroWordNet and the Referentie Bestand Nederlands (RBN). This creates a large semantic database which can be used for many natural language processing technologies. We have used the lexical part of the database, which is provided by the RBN. This is a collection of many different types of words and their relative functions in the Dutch language, including the classification of their semantic types. Most of our manually classified words were similarly classified by Cornetto, but a handful of them were differently classified. This is because the RBN corpus uses more extensive semantic types and dimensions. A word such as *artist* that we would manually classify as concrete would be classified by Cornetto as "HUMAN, ANIMATE". Cornetto also placed words such as *morning* and *evening* into a "TIME" dimension. However, we defined these "TIME" words as abstract. We have tried to generalize the Animate and Time objects as respectively concrete and abstract words, but for example a word as *verdachte* (suspect) was classified as "HUMAN" by Cornetto, but classified by us as an abstract word. Other words that we had classified as concrete, but turned up as abstract by Cornetto were: *gram* (gramme), *lied* (song), *prooi* (prey), *roman* (novel) and *woord* (word). One can discuss whether these words are actually more often used as concrete or abstract in daily language. This is why the decision was made to keep the manually selected list as the basis for the word list. However, we used the Cornetto results as a backup reference to compare these with the manually classified words. Apart from the aforementioned words, no large discrepancies were found.

Ambiguity

It is impossible to make a clear "black and white" distinction about whether words are abstract or concrete. Every word we say or write derives its meaning from the context it is used in; even a very abstract word such as *love* can be used in a concrete way like "Tonight, me and my *love* will have a romantic dinner". In the Cornetto database, the word *prooi* (prey) was seen as both concrete as abstract. We have kept the word concrete, even though an abstract way of using the word is very likely as well. This is something we had to take into consideration while preparing our data. However, to derive statistical information about the data we have gathered, we will keep our strict abstract and concrete distinctions because we believe that this can give us useful information about the sort of answers the children give.

3.1.5 Relationship Templates

Currently, there are 18 different templates in which users can add data to OMCS Dutch. For our experiment, we have selected those templates from OMCS that children can immediately comprehend, without making too many grammatical errors. Next to that, the templates should be used to describe noun phrases (NP), so templates in which you have to describe events by adding verb phrases (VP) such as "The last thing you do when you _(VP)_ is _(VP)_ ." were not usable for our experiment. We have made a selection of six relationship templates which you can see in Table 3.2. The column "OMCS Relation" is the same in every language. In OMCS English, these six relations contain over 70% of all the assertions that have been added, with UsedFor and CapableOf together covering around 40% of all assertions [Liu and Singh, 2004].

Dutch Template	**English translation**	**OMCS relation**
is gemaakt van	is made of	MadeOf
is een soort van	is a sort of	IsA
kan gebruikt worden voor	can be used for	UsedFor
kan	can	CapableOf
is een deel van	is a part of	PartOf
vind je vaak dichtbij	you are likely to find ___ near ___	AtLocation

Table 3.2: Selection of OMCS relationship templates which the children could choose for their descriptions.

3.2 Pilot Study

A short pilot experiment was held at a daycare center. This provided us with some useful information. Apart from this being a test to check any technical issues, there also was the question if children were indeed enthusiastic about playing it. Another issue was the reliability of the answers if the experiment was to be held online without any supervision. In the pilot, six children, aged eight to twelve years old played the game online, but with supervision. While the game was played, they were allowed to ask the supervisors any questions about the gameplay, but not about any of the words or answers. The eight year old children took a relatively long while (more than 30 minutes) to finish the game. Most of their answers were correct, even though very general and repetitive at times. For example, they would answer an abstract word such as "year" with "a *year* is made of *the year is over*" and "a *year* is a *year*". Another problem with the younger children was their spelling, some words were difficult to recognize. The older children answered with grammatically correct answers and their spelling was obviously better as well. When the children were asked if they would want to play the game again, all of them said they would.

The children in the pilot study were presented with an overall equal amount of words split up into concrete, abstract and difficulty level. However, results from the pilot study showed that the children were hesitating a lot when they were presented with an abstract or difficult word; sometimes one word took them over five minutes to describe. Even though the children knew about the ability to skip a word without any consequences, they seemed really determined to describe that word. However in some cases, the concentration declined so much that they started chatting about other stuff to the supervisors. The children should be kept stimulated to play the game, that is why the amount of easy, concrete words has been raised in the list of words to 40 words instead of 20. This will increase the chance that children be presented with a concrete word, which will keep them enthusiastic at the start of the game. Before the experiment began, one word, "gas", with a GM of 58 was initially sorted into the average category, but later placed to the easy category because average words were actually words between 48 and 58. This is why the distribution is not exactly equal for average and difficult. This difference can be rectified by using the percentages as a reference for the distribution of the given answers.

Difficulty	**Concrete**	**Abstract**	**Total**	**Percentage**
Easy	41	20	61	43.6%
Average	19	20	39	27.8%
Difficult	20	20	40	28.6%
Total	80	60	140	100%
Percentage	57.1%	42.9%	100%	

Table 3.3: Distribution of assigned words in concrete/abstract and level of difficulty.

Table 3.3 shows the relative and absolute distribution of the words put in the database. These were randomly assigned to the children. There are around 14% more easy words than average and difficult words and over 14% more concrete words than abstract words. If the children would answer all the assigned words without skipping any of them, the distribution of the answers should be the same.

3.3 Gameplay

When a child visits the web site he or she is first presented with a small introduction about the purpose of the game; a fictional robot named "Rob" wants to be taught common sense, and tells the children that he needs their help to teach him that. The child is then asked to play two word games. On the next page he/she is asked to enter some basic demographic information such as age, grade and gender.

After this is entered, the child is shown a page with an explanation by Rob about the Describer game, see Figure 3.1. We have chosen to let the participant be the "Describer" first, which means he or she has to describe an assigned word. The word would be randomly picked from the database and would then be shown on screen. The child has to describe three words from each level of difficulty, going from the easy level to the difficult level. See Figure 3.2 for a screenshot. If the child cannot come up with a description of a difficult word he/she will get another difficult word to describe, until three words of that difficulty are described. The child can then pick three of the six relationship templates to describe that assigned word. It is impossible to use the same relationship more than once when describing a word. The child also has the option to skip a word when he or she could not think of any fitting description.

Verbosity has to rely on two players playing one game at a time, with each of the participants playing either the "Narrator" (here: Describer) or the "Guesser". Our game is a single-player game, because we are only interested in the children's descriptions and do not want to game be dependent on a second participant.

After the child has finished describing nine different words, he or she is asked to guess nine randomly assigned words of which only the descriptions are given. These descriptions consist of the descriptions that other children have given in their "Describer" role. Rob will also give a short explanation on how this game works, see Figure 3.3. If a word is correctly guessed, Rob will congratulate the child. In Verbosity, the "Guesser" part serves as a guide to monitor if the given descriptions make any sense.

Figure 3.1: Short explanation by "Rob the Robot" about the Describer game.

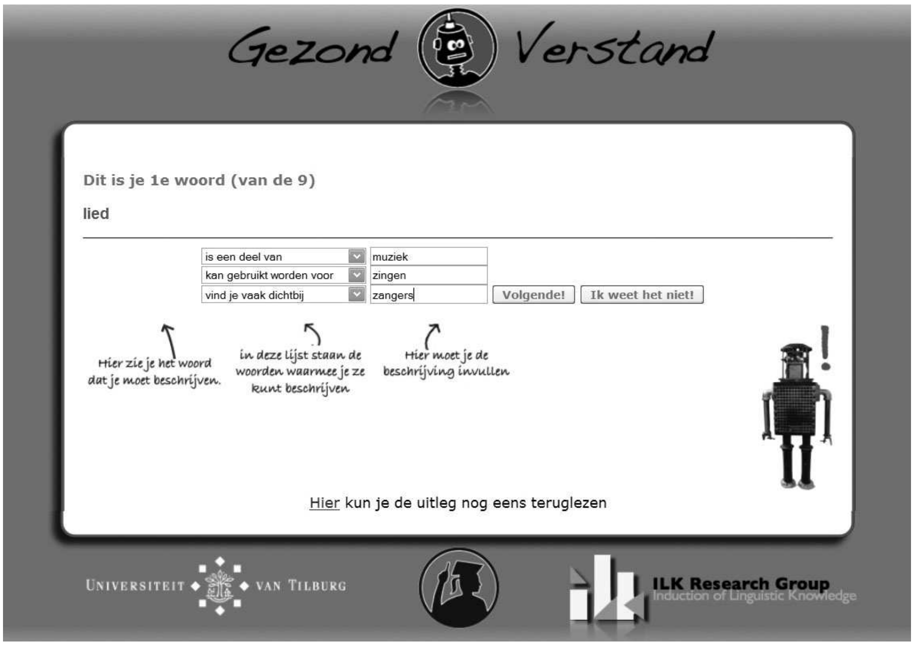


Figure 3.2: Example of the Describer game.

Figure 3.3: Short explanation by "Rob the Robot" about the Guesser game.

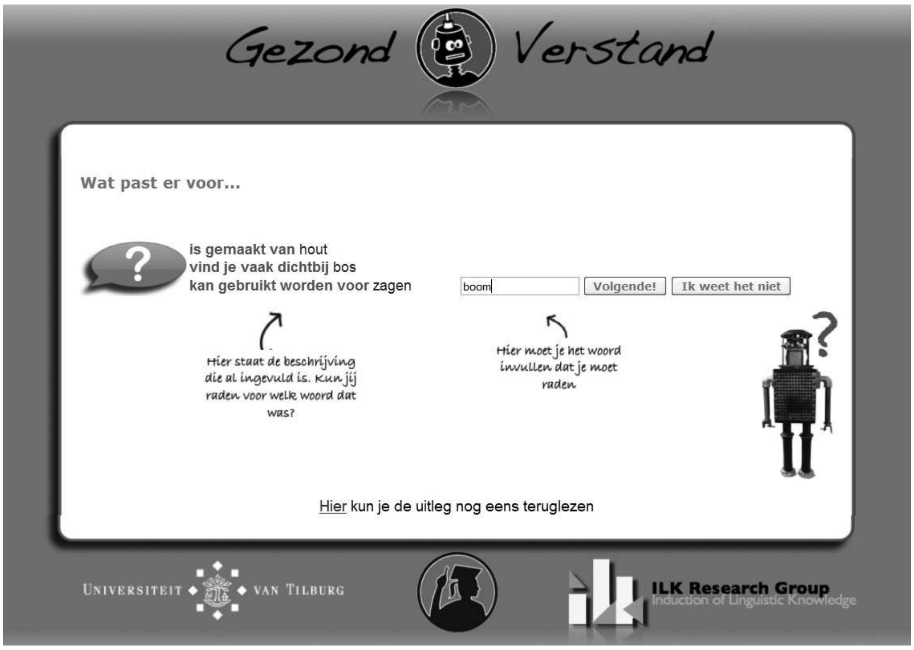


Figure 3.4: Example of the Guesser game.

In our research the focus lies on the kinds of descriptions the children give, because their answers will be manually checked anyway. However, in our pilot study, many children did enjoy the Guesser role very much, especially when they got some feedback about points and high scores. This helps getting the children enthusiastic about playing the game. But as mentioned in Section 2.3.3, it is difficult to emulate a Guesser. This is also why we have not used any of the children's guesses for our experiment results.

3.4 Subjects

For the experiment we have invited participants of the "Kinderuniversiteit" (Children's university) to visit our website with the game. The "Kinderuniversiteit" is a series of lectures for children who are in the upper grades of primary school. In these lectures many different scientific subjects are explained in a fun way to the children. These lectures are given by professors from Tilburg University and Eindhoven University of Technology.

The children who attend the "Kinderuniversiteit" vary from nine to twelve years old. Our experimental subjects are 10 to 12-year-old children, because at the age of nine, spelling and vocabulary knowledge is not adequate enough to describe the difficult words that were shown to them. Furthermore, results from our pilot show that the spelling of 10-year-old children is considerably better. Next to that, we also need the subjects to still be at primary school. This is because the Schrooten and Vermeer corpus we have used is based on words that are presented to the children during their years at primary school. Also, because ten is an age in which children consciously start to make use of words and symbols (see also Section 2.2), it will be interesting to see what kind of descriptions they give and how reliable they are. The children were invited to participate in the experiment by sending their parents a letter in which they were asked to visit the website where the game could be played. These invitations were send out at the beginning of the summer holiday, so many children had ample time to visit the website. Data was collected during a three-week period.

Vocabulary Knowledge

Age	**Amount of Words**
9 years old	7,000 words
10 years old	10,000 words
11 years old	13,000 words
12 years old	16,000 words

Table 3.4: Estimated amount of receptive Dutch vocabulary of native Dutch speaking children in primary school. Source: [Van de Guchte and Vermeer, 2003]

The vocabulary size of our subjects is already relatively extensive at their age; among 10-year-old children with Dutch as their first language, the size is around 10,000 words, among 12-year-old children this is already 16,000 words. For children who do not have Dutch as their first language at home, this amount is considerably lower with 9,800 words for 12-year-old children. Table 3.4 illustrates the growth of the vocabulary for children with Dutch as their first language. We expect that only a marginal amount of the subjects will have Dutch as a second language at home. This is why we have only taken the vocabulary knowledge of native Dutch speaking children into account.

Chapter 4

Results

4.1 Basic Frequencies

Within three weeks we had collected answers from more than 200 played games. After cleaning up false entries, there were 151 correctly played games left, played by 123 unique persons. Of the 123 subjects 60 were female and 63 were male. When the describing role was completed, each subject would have given 27 answers (nine assigned words that needed three different descriptions). A total of 4,077 descriptions were given. In table 4.1 an overview is given of the amount of words answered per gender and their percentages.

Gender	**Unique Persons**	**Percentage Persons**	**Frequency Descriptions**	**Percentage Words**
Male	63	51.2%	2,106	51.7%
Female	60	48.8%	1,971	48.3%
Total	123	100%	4,077	100%

Table 4.1: Overview of unique players and total amount of descriptions distributed by gender.

Of the 123 subjects, 103 children finished one game and 20 children played the game more than once. One of the kids seemed to enjoy it so much that she played it eight times. Playing the game more than once was allowed, because the more data is collected, the more results can be derived. Only one of the subjects did not have Dutch as a first language at home. This means that, as expected, we did not have to take any language deficiencies into account in the results.

The average age was 10.9 years old; 52 children were 10 years old, 36 children were 11 and 35 were 12 years old. 76 children were in ”groep acht” (sixth grade), 47 were in ”groep zeven” (fifth grade). Since this test was taken in the summer holiday, and most children were switching from one grade to another, we cannot say much about the correctness of the grades they were in at the moment of the experiment. This is why the children’s grade has not been used for further analysis.

4.2 Assigned Words - Frequencies

In this section we will present frequencies of the assigned words that were described by the children. The word difficulty level had an equal distribution, because each child had to answer three words of each level. However, certain words have been answered considerably more often and others considerably less than others. This can not be 100% attributed to the database, because when randomly presenting words, the amount of presented words should be more or less equal. However, we cannot draw any strong conclusions from this assumption, because the game did not save any skipped words. Theoretically speaking, if each of the 4,077 descriptions were described in an equal amount, each assigned word would have to be described 29 times. However, there are large differences in the amount of times a word was described.

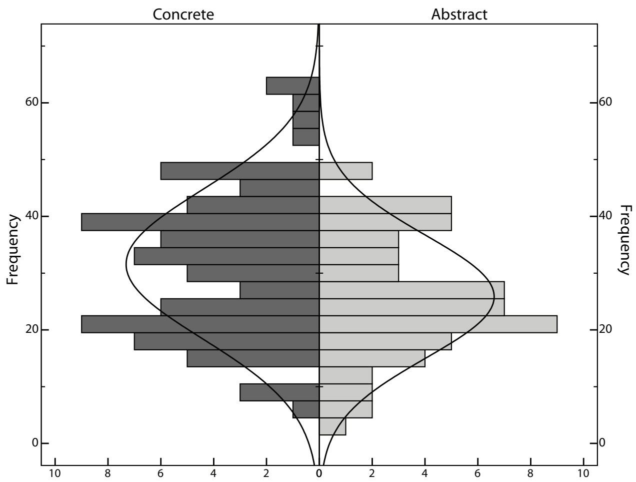


Figure 4.1: Distribution of frequency of Concrete and Abstract words.

Figure 4.1 shows the distribution of the concrete and abstract words. The "Frequency" axis shows how many times a word was described. The Concrete and Abstract columns show how many words have had the same amount of descriptions. For example: there was only one abstract word that was described three times, but there were eight abstract words that were described 30 times. The abstract words are more present in the lower frequencies, and concrete words are more present in the higher frequencies. This means that the words with the most descriptions are more often concrete and the words with the least descriptions are more often abstract. Words that were only described a few times were *avond (three times)* ("evening"), *opzet (six times)* ("intent, design"), *raam (six times)* ("window") and *vraag (six times)* ("question"). Words that were described very often were *cd (63 times)* ("cd"), *bekeuring (63 times)* ("fine ticket"), *gram (60 times)* ("gramme") and *gewei (54 times)* ("antler"). These frequencies show that there could be a difference in the amount of times a word was actually presented to the children by the game: "raam" is a word that should be relatively easy to describe, while on the other hand "bekeuring" seems to be a word that not everyone would find easy to describe. There could be two explanations for this; the first is that the game only presented "raam" a handful of times to the children, and the second is that children actually skipped "raam" often.

4.3 Assigned Concreteness - Frequencies

Table 4.2 shows the relative and absolute amount of descriptions of concrete and abstract words compared with the relative amount of concreteness in assigned words. Of the described words 61.1% were concrete words and 38.9% were abstract, which means that there is a tendency towards describing concrete words more often (see also Section 3.3).

	Frequency	**Percent**	**Assigned Percentage**
Concrete	2,493	61.1%	56.8%
Abstract	1,584	38.9%	43.1%
Total	4,077	100.0%	100.0%

Table 4.2: Total and relative amount of concrete/abstract words described by the children vs. relative amount of concrete/abstract words presented to the children.

4.4 Relation Templates - Frequencies

Table 4.3 shows which templates are chosen for the descriptions. The highest frequency is *IsA*, with 22% of the total amount of descriptions. The template which was chosen the least is *PartOf* with 12.8% of the total amount of descriptions. The distribution of these templates is relatively equally distributed, even though the children were free to choose which template to use.

Template	**OMCS relation**	**Frequency**	**Percent**
vind je vaak dichtbij	AtLocation	635	15.6 %
kan	CapableOF	703	17.2 %
is een soort van	IsA	898	22.0 %
is gemaakt van	MadeOF	588	14.4 %
is een deel van	PartOf	522	12.8 %
kan gebruikt worden voor	UsedFor	731	17.9 %
Total		4,077	100 %

Table 4.3: Overview of templates chosen for the descriptions.

4.5 Causal Links between Gender and Age

This section will present whether any causal links are found between demographic factors gender and age with concreteness, template preference and description length. In other words, we want to see if the descriptions from boys are different from those of girls, and whether there is a difference in the descriptions between 10, 11 and 12-year-old children.

4.5.1 Causal links between Gender

Concrete/Abstract

First, a gender comparison between the amount of described concrete and abstract words is made. Of the 2,106 male descriptions, 60.8% is concrete and 39,2% is abstract. Of the 1,971 female descriptions, 61.5% is concrete and 38.5% is abstract. We can say that there is hardly a difference between boys and girls in their choice of describing a concrete or abstract word.

Template Preference

The distribution of the relationship templates chosen by children to describe assigned words also shows no causal link between gender either.

Description Length

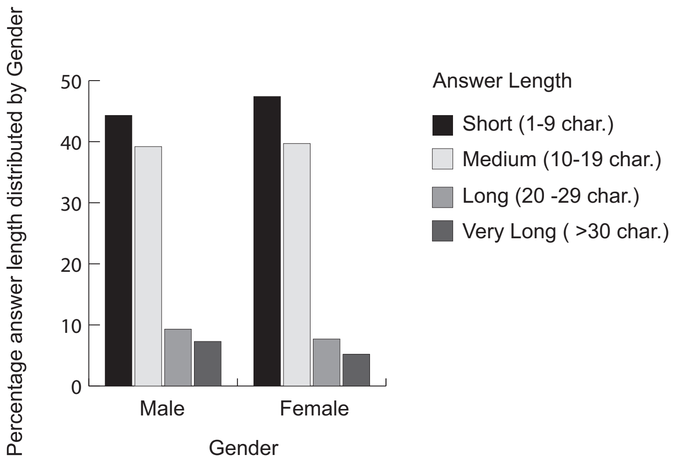


Figure 4.2: Length of descriptions given distributed by gender

Figure 4.2 shows the frequency of lengths of descriptions that were given. The lengths of the answers are measured by counting the total amount of characters used in the description (including spaces and punctuation). These measurements are divided into four categories: short (1 - 9 characters, around one or two words), medium (10 - 19 characters, around two to four words), long (20 - 29 characters, around four to six words) and very long (more than 30 characters, more than around five words). Only a small amount of the length of the descriptions is long and very long and only a small difference in gender can be seen: boys give a slightly larger amount of longer (long and very long) answers than girls. However, these longer answers only account for a small portion of the total amount of answers.

These three results show that differences between gender are marginal. For our further analysis, we will therefore not make any other comparisons between gender.

4.5.2 Causal Links between Age

Table 4.4 shows the distribution of age in the descriptions that are given. 43.7% of the descriptions are given by 10-year-old children, 27.2% by 11-year-old children and 29.1% by 12-year-old children.

Age	**Frequency**	**Percent**
10	1,782	43.7%
11	1,107	27.2%
12	1,188	29.1%
Total	4,077	100.0%

Table 4.4: Amount of descriptions given, distributed by age.

Concrete/Abstract

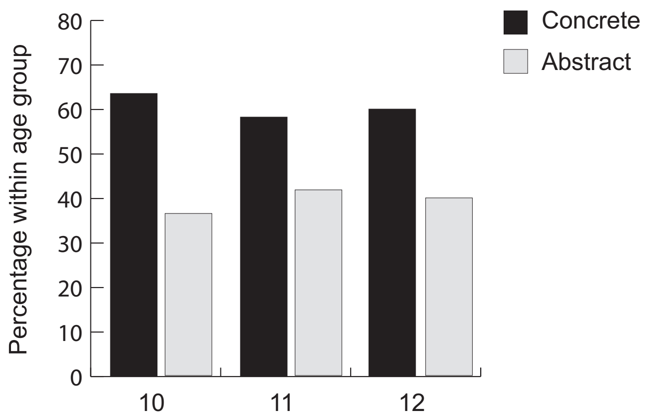


Figure 4.3: Relative amount of concrete abstract words described, distributed by age.

Figure 4.3 shows the relative amount of concrete and abstract words that are described by the children, distributed by age. The difference between number of described abstract and concrete words is largest among 10-year-old children. This shows that 10-year-old children choose to describe more concrete words and skipped more abstract words. Among the older children, the distribution of concrete and abstract words becomes more similar to the distribution that in which the words were distributed initially (see Table 3.3). This means that the 11 and 12-year-old children skipped less abstract words.

Template Preference

No large differences are found between age and whether there is a preference for a certain template. Two small differences are found, but these may be coincidental. The first difference is that the older the age, the more they prefer *MadeOf*. Secondly, 11-year-old children choose *PartOf* more often than the other age groups. However, these differences are marginal.

Description Length

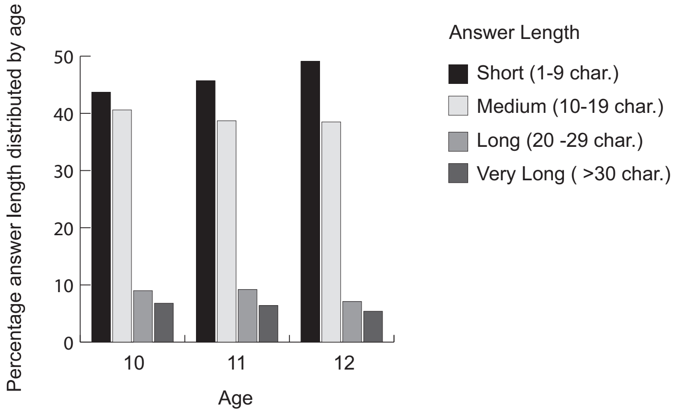


Figure 4.4: Relative amount of concrete abstract words described, distributed by age.

Figure 4.4 shows the relative amount of length of the descriptions that the children have given. What we can see is that 10-year-old children seem to give an almost equal amount of short and medium descriptions. Also, the higher the age, the greater the relative amount of short descriptions given. For longer answers we can say that 12-year-old children seem to give less long and very long answers to describe a word. However, longer answers account for only a small percentage of the total amount of descriptions given.

Overall we can say that the basic frequencies of the demographic characteristics from the participants are very similar. Differences that are found are minimal. This is why we have chosen to use the entire population to answer our main questions.

4.6 Correctness of the Descriptions

We manually examined a random selection of 25% of the 4,077 descriptions that are given. The descriptions are distributed into four different categories:

- Semantically incorrect answer or nonsense filled in.
- Poor, not semantically incorrect but does not give much information.
- Correct, but with a typographical error.
- Correct.

The percentages and counts of the examined descriptions are shown in Table 4.5. By far most of the answers are correct; 92% is correct, plus 3.6% that is also correct but with a typo. This means that when you combine these two categories, 95.6% of all descriptions are correct. Only 1.2% is incorrect, an incorrect example is "*tonic is made of fish*" and 3.2% of all answers is marked "poor", for example: "*a bag is part of life*".

Status	Frequency	Percent
Wrong	13	1.3%
Poor	32	3.1%
Correct-Typo	37	3.6%
Correct	943	92%
Total	1,025	100.0%

Table 4.5: Frequency of manually examined descriptions, taken from a 25% random selection of the descriptions.

When these results are split at concreteness and we look at the relative amount of the status for abstract and concrete words, it can be seen in Table 4.6 that "Abstract" scores lower in "Correct" compared to "Concrete". However, "Abstract" has a higher amount of correct descriptions with a typo, which makes the total of correct words in abstract (90.0% + 4.8% =) 94.8% and for concrete words (93.4% + 2.8 % =) 96.2%.

Status	Concrete	Abstract
Wrong	1.3%	1.2%
Poor	2.5%	4.0%
Correct-Typo	2.8%	4.8%
Correct	93.4%	90.0%
Total	100.0%	100.0%

Table 4.6: Relative amount of correctness in concrete/abstract

We also checked whether there are differences in correctness compared with the difficulty level of the assigned words. The relative amount of correctness of each level of difficulty is shown in Table 4.7. Difficult scores over 3% lower at "Correct" than the other two difficulty levels. "Difficult" scores higher in the "Correct-Typo" status, which means that more typos are made in the "Difficult" level.

Status	Easy	Average	Difficult
Wrong	1.5%	0.3%	2.0%
Poor	2.4%	3.1%	3.8%
Correct-Typo	2.7%	3.4%	4.6%
Correct	93.3%	93.1%	89.6%
Total	100.0%	100.0%	100.0%

Table 4.7: Relative amount of correctness in Difficulty Level of assigned words.

4.7 Vocabulary Level of the Children's Descriptions

In order to find out what the vocabulary level of the children's descriptions is, the Schrooten & Vermeer corpus [Schrooten and Vermeer, 1994] (further referred to as "SV corpus") will be used. This time the descriptions are matched with the words in the SV corpus. The preparation of this data will be explained in the next section.

4.7.1 Preparing the Data

The children's descriptions are cleaned up by removing all punctuation and extra spaces. Next, the descriptions are combined into complete sentences with their assigned words and templates. This is necessary because the next step is automatically lemmatizing and tagging the sentences by using Tadpole (which needs complete sentences to be able to automatically tag and lemmatize the words). Tadpole is a memory-based tagger and parser for Dutch, which uses k-nearest neighbor classification [van den Bosch et al., 2007]. The Tadpole output, consisting of split-up lemmatized sentences, is then combined into complete (grammatically incorrect) sentences (by using a script) to match them with the player id's of the children in the database. The complete sentences have to be split up again by removing the assigned words and relation templates, because these are essentially not part of the words that the children had entered. This leaves us with lemmatized versions of the children's descriptions. Next, function words are deleted from the lemmatized descriptions by using another script. The remainder of the words are then separated into single words again, but now connected to their player id's. The single words mainly consist of verbs, nouns and adjectives. These are all used to measure the difficulty level of words given by the children.

To avoid ambiguity in the SV corpus, many lemmas contain an added token of semantical meaning consisting of an underscore and a semantic meaning such as *drinken_N drinken_V* (English: *drink_N* and *drink_V*) to point out the difference between the noun and verb in this case. To be able to match the lemmatized descriptions with the SV words, a new column with a cleaned up version of the lemmas in the corpus has to be created. After creating a new table of the SV corpus in the database, the single words of the children are then ready to be matched to their equals in the SV corpus. The result is a match with the clean words in combination with all their semantic versions in the SV corpus. However, the actual unambiguous meaning of the word used in the description still has to be manually selected by looking at the context of the word in the original description given by the children.

After removing the redundant semantical meanings, a total of 5,060 words matched with their SV corpus-equals. There were 1,485 words that did not match with the SV corpus. These were usually words that had a typo, composed words or simply words that were not recognized by Tadpole and because of that not lemmatized.

The geometric mean of the SV lemmas are sorted into the same difficulty levels as used for the assigned words that the children had to describe. See Section 3.1.3 for these details. The next section will deal with the results that originated from this match.

4.7.2 Results

Table 4.8 shows the counts and relative results of the lemmatized descriptions of the children that match with their equals in the SV corpus. A total of 6,545 words were prepared to be matched, 5,060 descriptions actually matched. This is 77.3% of the total. Of all the matched 5,060 words, 74.8% is defined as easy, 16.6% is defined as average and 8.6% as difficult.

Difficulty Level	Frequency	Percent
Easy	3,789	74.8%
Average	839	16.6%
Difficult	435	8.6%
Total	5,060	100.0%

Table 4.8: Frequency of difficulty level used in the descriptions in combination with the difficulty level of assigned words.

Assigned words vs. Description Difficulty Level

Several statistical measures are conducted to find out if difficult assigned words are more often described with difficult words, and if easy words are more often described with easy words. In Figure 4.5 the counts of the assigned words levels are shown in combination with the difficulty levels of the descriptions. Most of the words that are used in the descriptions are easy. This is illustrated by the left "easy" bar, which has a much higher frequency than the left bars of the other two description levels. In the "easy" bar, the number of assigned difficult words is lower than assigned easy words. In the right "difficult" bar, the number of assigned difficult words is higher than the number of assigned easy words.

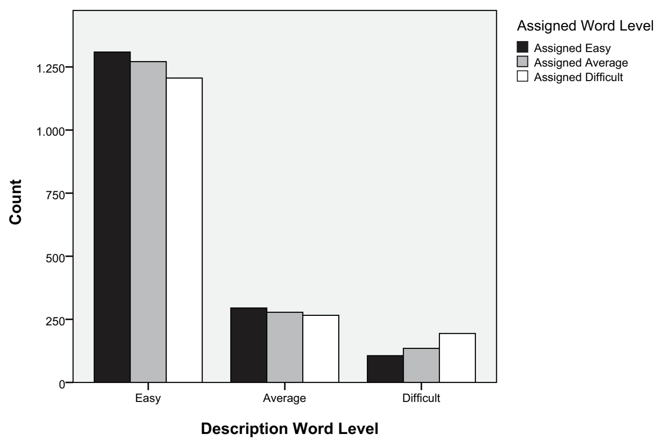


Figure 4.5: Difficulty level of descriptions in combination with given word levels.

A chi-square test (Cramer's V) is conducted to see if there is an equal distribution between the variables. The results from this test are: $\chi^2(4) = 33.04$, $p<.001$. This means that the chance that these difficulty levels are equally distributed is almost zero; there must be a significant relationship amongst the variables, which most likely seems to be between the easy and difficult levels when looking at Figure 4.5. Average word levels do not seem to show any large differences with the other levels.

To see if there is a significant effect between the easy and difficult levels, we conducted a follow-up chi-square (Phi) test between Given Easy/Difficult and Described Easy/Difficult. The results from this test are: $\chi^2(1) = 29.95$, p<.001. The effect size is .103. This means that there is a weak correlation between the word levels. This is also illustrated in Figure 4.6. The follow-up test between Assigned Easy/Difficult and Described Easy/Difficult found that assigned easy words are more likely to be described by easy words and assigned difficult words are more likely to be answered with difficult words.

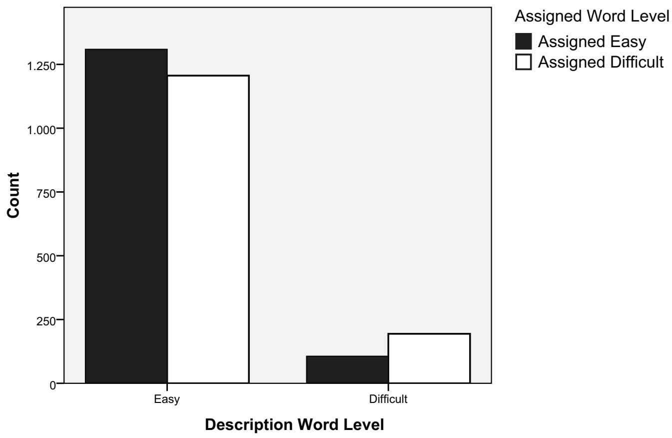


Figure 4.6: Described easy and difficult words in combination with assigned easy and difficult words.

Relation Templates vs. Description Difficulty Level

The relationship templates that the children chose to describe an assigned word with, are combined with the level of difficulty of the children's descriptions. This comparison shows whether a chosen template will often get easier or more difficult descriptions than other templates. Figure 4.7 gives an overview of the distribution of the difficulty level of descriptions that children have given in combination with the templates that children chose.

By far most words used for the descriptions are easy; this makes sense because most of the words people use in daily life are easy words (such as in the SV corpus). In the average and difficult words, the relative amount of the relationship frequencies are very similar. These frequencies can be related to the children's overall choice of relations shown in Table 4.3; those top frequencies are the same as in Figure 4.7. However, when comparing the relative amounts of chosen relationships in the easy words, *CapableOf* and *UsedFor* score much higher than in the other two difficulty levels. This means that most of the easy descriptions were used to answer the *CapableOf* and *UsedFor* relations. This difference may well be caused by the fact that in the *CapableOf* and *UsedFor* relations a Verb Phrase needs to be filled in. Verbs are usually easier words, because they are used more frequently

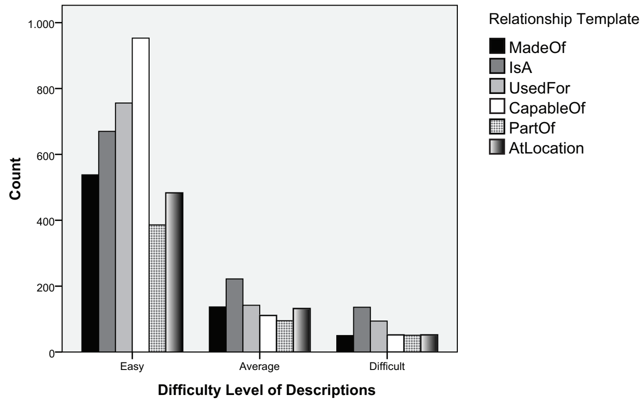


Figure 4.7: Difficulty Level of Descriptions in combination with chosen relation templates.

than nouns. This can also be checked in the SV corpus; verbs are often present in the words with a higher GM (which are the easier words). In the other relations, noun phrases need to be filled in.

4.8 Concreteness of the Children's Descriptions

Words used in the descriptions of the children are also split up into abstract or concrete. This is done to compare the assigned concrete and abstract words with the descriptions these words have received. The preparation of this data is explained in the next section.

4.8.1 Preparing the data

Tadpole's automatically lemmatized and tagged nouns of the descriptions are used in this comparison. Only nouns are used because the assigned words are nouns as well and because nouns are the most informative words in a description.

First, the lemmas that have been tagged "Noun" by Tadpole are selected, next any duplicates are removed. These unique lemmas are automatically classified by using the Cornetto Database (see also Section 3.1.4). Subsequently, the Cornetto results are manually converted to either an abstract or concrete value. This has to be done manually because Cornetto classifies words into more dimensions than abstract and concrete alone. This is also a good way to monitor the Cornetto classifications.

The unique list of lemmas is then put into the database. Next, this list is matched against all the nouns that were given by the children. The results of this are dealt with in the next section.

4.8.2 Results

A total of 4,061 (non-unique) lemmatized nouns given by the children matched with the classified words of Cornetto. These frequencies are shown in Table 4.9. There are 191 missing values. This is caused by the fact that Tadpole misrecognized some words as nouns. Most of these errors are caused by typos, some others by grammatically incorrectness of the composed sentence. Of the 3,870 words that did match, 67.3% of the described nouns are concrete and 32.7% are abstract.

Concrete/Abstract	**Frequency**	**Percent**
Concrete	2,606	67.3%
Abstract	1,264	32.7%
Total	3,870	100.0%

Table 4.9: Frequency of assigned concrete/abstract words in combination with concreteness found in the descriptions.

Concrete/Abstract Descriptions vs. Assigned Concrete/Abstract Words

To answer the question whether assigned abstract words are described with abstract descriptions and assigned concrete words by concrete descriptions, a chi-square test (Phi) has been conducted. The results from this test are: $\chi^2(1) = 7.44$, p<.001. The effect size is .438. This means that there is indeed a strong correlation between the assigned concrete/abstract words and the concreteness of the descriptions.

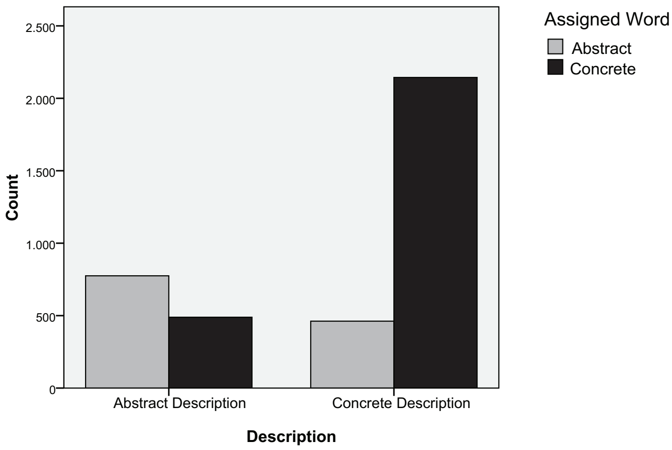


Figure 4.8: Concreteness found in the descriptions in combination with assigned concrete/abstract words.

Figure 4.8 shows that abstract words will more often be described with abstract words, and that concrete words will usually be described with concrete words. The figure also shows that the two bars in the left abstract description are much closer together than the two bars in the concrete description on the right. This is because compared to concrete words, a smaller percentage of abstract words was shown to the children, which consequently yields a smaller amount of abstract descriptions. The large "assigned" concrete bar in the concrete description on the right is caused by the fact that more concrete words were shown to the children and assigned concrete words usually receive more concrete descriptions.

Concreteness of Descriptions vs. Assigned Difficulty Level

Figure 4.9 shows the distribution of the difficulty level of assigned words with the abstract- or concreteness of descriptions the children have given. The difficulty levels of assigned words are equally distributed; every child is presented with three words of each level. The bar chart in Figure 4.9 makes clear that most of the words children give are concrete words. What is interesting is that abstract descriptions are used to describe difficult words most of the time, and concrete descriptions are used more often to describe easy assigned words.

To see if there is a significant effect between difficulty level of the assigned words and the abstract- or concreteness of the descriptions, we conducted a chi-square (Cramer's) test. The results from this test are: $\chi^2(2) = 22.90$, p<.001. The effect size is .077. This means that there is indeed a weak correlation between the amount of concrete and abstract words used in the descriptions in combination with the assigned difficulty level. Abstract descriptions are more used to describe difficult words and concrete descriptions are used more often when describing easy words.

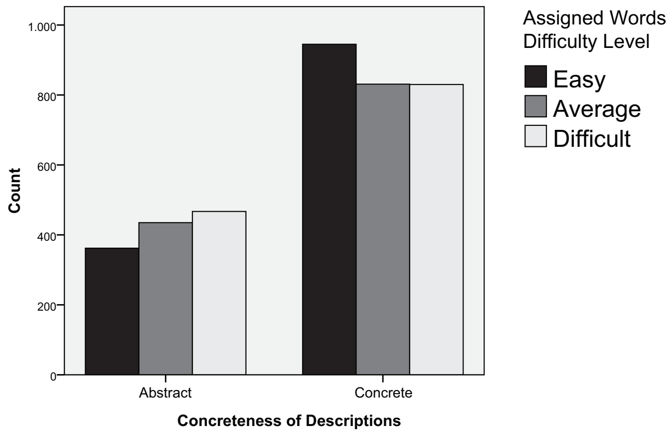


Figure 4.9: Concreteness of description in combination with distribution of difficulty level of assigned words.

Concreteness of Descriptions vs. Relation Templates

Another aspect we want to show is which abstract and concrete descriptions are used for the relation templates. These results are shown in Figure 4.10. In the abstract and concrete descriptions, two relations have completely different results: *MadeOf* and *AtLocation* are low in the abstract descriptions and have the highest frequencies in the concrete descriptions. In other words, abstract descriptions are hardly used in the *MadeOf* and *AtLocation* relations, and concrete words are used most in the *MadeOf* and *AtLocation* relations. This difference can be caused by the fact that you hardly need to use any abstract words to describe a material (in *MadeOf*) or a location (in *AtLocation*), unlike

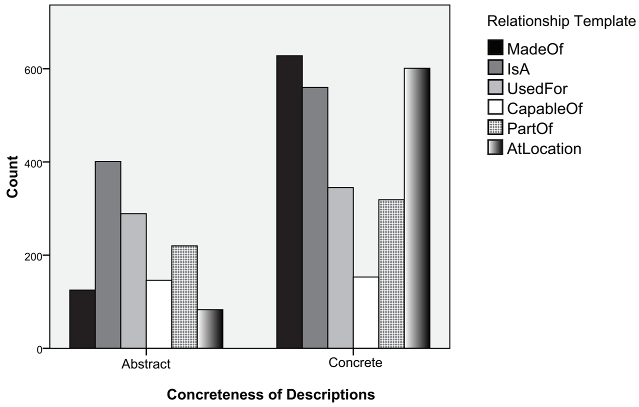


Figure 4.10: Concreteness of descriptions in combination with chosen relation templates.

concrete words, where it is basically necessary to use them for describing materials and locations.

Chapter 5

Conclusions and Further Research

5.1 Conclusions

Children aged ten to twelve have built up a decent enough vocabulary basis (10,000 to 16,000 words) to be able to describe a word. Next to this, most of our participants are in the Concrete Operational stage. This means that they are more engaged in here-and-now situations and more with concrete facts, this is beneficial for describing (especially concrete) nouns: it is in the children's nature to think about facts in that way. 10-year-old have indeed described more concrete and less abstract nouns than the older children. This difference may be caused by the fact that 11 and 12-year-old children start to learn to think more abstractly according to Piaget.

A total of 123 children have played a complete game in a period of three weeks time. Dozens of children have played the game more than once, so playing a simple game seems to be much fun for children.

The children's descriptions were very reliable; over 95% of the given descriptions that were manually checked are correct. The amount of spelling errors and typographical errors is also very low; less than 4% of the correct answers contained a typographical- or spelling error. Concrete words and easy words score a little bit better in the amount of correct descriptions, however, the differences in scores are marginal. The difficulty level and concreteness of the words hardly make any difference in correctness of the description. This means that children can play this game with any kind of word from the Schrooten and Vermeer corpus, and they will describe the word correctly most of the times.

The language use in the children's descriptions also give some interesting results. Overall, most of the words the children use are easy and concrete. This makes sense, because most of the time in daily life, we also use easier and more concrete words. This is also shown in the SV corpus where words with a high frequency are words that are often used. Difficult assigned words are described more often with difficult words in the descriptions, the same counts for assigned abstract words, these are also described more often with abstract words.

It is obvious that the children enjoy playing the game, probably even more than adults would, because adults can become bored more quickly with such a simple game, and to adults, making commonsense knowledge explicit seems to come harder. This study shows that children are indeed a reliable source for adding commonsense knowledge to the OMCS database.

5.2 Further research

This experiment aimed at finding out whether children can be a reliable source for adding commonsense information for the OMCS database. Results show that children are indeed a reliable source. If a Verbosity-like game were to be combined with the OMCS-Dutch database, much knowledge could be gathered, but before doing so, it can help to keep in mind certain findings from our research.

Abstract words are less often described than concrete words, and abstract words elicit a slightly larger amount of incorrect descriptions, because of this, the distribution of concreteness of the assigned words in the game should be thought of. This does not mean that abstract words should not be used in the game, on the contrary, it is very important that abstract words will also receive enough descriptions.

In our experiment we did not utilize the reliability factor of the "Guesser" part as Verbosity has. In Verbosity, only words that are correctly guessed by the Guesser are saved and added to the commonsense database. This means that there is almost no need to monitor the results. If the guesser part is also implemented into our game, it is likely that manual monitoring would spot few errors. However, implementing a guesser-option creates two possibilities: either the game would become dependent on two players or if the game continues to be single-player, a list of related Dutch words should be created first to let the computer guess words like a human would. The level of spelling of the (human) guesser is also an important factor; maybe an automatic spelling corrector of the guessed words would have to be used, because certain words can be difficult to spell for children. This could prevent situations in which a child actually knows what word is asked but misspells it and the word would be seen as incorrect.

Also, if the OMCS-Dutch corpus can become large enough without the use of our game, the children's descriptions can be compared to those of the OMCS-Dutch participants. Then it would be possible to see in what way children's descriptions differ from those given in OMCS.

In OMCS, participants have to think of their own concepts to describe; in our game, words are assigned to people. A positive side-effect from this is that people will try to describe words which they may never would have thought of themselves. The disadvantage of this strategy is that we should put some thought in the choice of assigned words, as this is an important bias in the outcomes. If this game were to be used next to the OMCS-Dutch website, a varied and broad collection of commonsense knowledge could be gathered.

Bibliography

[Ackermann, 2001] Ackermann, E. (2001). Piaget's Constructivism, Papert's Constructionism: What's the difference? *Future of Learning Group Publication.*

[Atherton, 2008] Atherton, J. S. (2008). Learning and teaching: Piaget's developmental theory. `http://www.learningandteaching.info/learning/piaget.htm` Retrieved: November 15th, 2008.

[Bhattacharya, 2008] Bhattacharya, L. & Han, S. (2008). Piaget and cognitive development. in m. orey (ed.), emerging perspectives on learning, teaching, and technology. `http://projects.coe.uga.edu/epltt/` Retrieved: November 15th, 2008.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* The MIT Press.

[Field, 2003] Field, J. (2003). *Psycholinguistics: A Resource Book for Students.* Routledge.

[Havasi et al., 2007] 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.*

[Landauer, 1986] Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, 10(4):477–493.

[Lenat, 1995] Lenat, D. B. (1995). Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38.

[Liu and Singh, 2004] Liu, H. and Singh, P. (2004). ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal)*, 22(4):211–226.

[McCarthy, 2007] McCarthy, J. (2007). From Here to Human-level AI. *Artifial Intelligence*, 171(18):1174–1182.

[Minsky, 1988] Minsky, M. (1988). *The Society Of Mind.* Simon & Schuster.

[Minsky, 2006] Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* Simon & Schuster.

[Molnar, 2005] Molnar, Andrew E. & Hertenstein, M. J. (2005). Piaget, Jean (1896–1980). *Encyclopedia of Human Development.* http://www.sage-ereference.com/humandevelopment/Article_n486.html Retrieved: November 16th, 2008.

[Mueller, 2003] Mueller, E. T. (2003). Thoughttreasure: A natural language/commonsense platform. http://alumni.media.mit.edu/~mueller/papers/tt.html.

[Papert, 1980] Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas.* Basic Books.

[Papert, 1999] Papert, S. (1999). Papert on piaget. *Time magazine: "The Century's Greatest Minds"*, page 105. http://www.papert.org/articles/Papertonpiaget.html Retrieved: November 21st, 2008.

[Roush, 2006] Roush, W. (2006). Marvin Minsky on Common Sense and Computers That Emote - Interview. http://www.technologyreview.com/computing/17164/ Retrieved: November 23rd, 2008.

[Schrooten and Vermeer, 1994] Schrooten, W. and Vermeer, A. (1994). Woorden in het Basisonderwijs. 15000 woorden aangeboden aan leerlingen. http://let.uvt.nl/general/people/avermeer/wrdlstin.htm.

[Singh, 2002] Singh, P. (2002). The Open Mind Common Sense Project, MIT Media Lab.

[Singh et al., 2002] Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. http://citeseer.ist.psu.edu/singh02open.html.

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

[Van de Guchte and Vermeer, 2003] Van de Guchte, C. and Vermeer, A. (2003). Een Passende Woordkeus: het kiezen van woorden voor woordenschatlessen. *Toegepaste Taalwetenschap*, 69(1):9–23.

[van den Bosch et al., 2007] van den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In Dirix, I. P., Schuurman, I., Vandeghinste, V., and Eynde, F. V., editors, *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114.

[von Ahn and Dabbish, 2004] von Ahn, L. and Dabbish, L. (2004). Labeling Images With A Computer Game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA. ACM Press.

[von Ahn et al., 2006] von Ahn, L., Kedia, M., and Blum, M. (2006). Verbosity: a game for collecting common-sense facts. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78, New York, NY, USA. ACM Press.

[Vossen et al., 2007] 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., editors, *Proceedings DIR 2007*, pages 89–96.

Samenvatting

In deze thesis wordt onderzocht of kinderen een betrouwbare bron zijn voor het toevoegen van Gezond Verstand kennis aan de Open Mind Common Sense database. Om deze kennis te verzamelen, hebben we een spel ontwikkeld dat gebaseerd is op een bestaand gezond verstand spel *Verbosity*, waarin spelers gevraagd wordt om verschillende zelfstandig naamwoorden te omschrijven en te raden. De zelfstandig naamwoorden zijn gehaald uit een corpus dat is gebaseerd op literatuur dat aan kinderen wordt getoond gedurende hun basisschool tijd. Deze woorden worden onderverdeeld in concrete en abstracte woorden en in verschillende moeilijkheidsniveau's. Aangezien het doel is om de verzamelde data toe te voegen aan de al bestaande Open Mind Commons Dutch website, worden de deelnemende kinderen gevraagd om tussen zes relatie-templates, gehaald uit de OMCS-Dutch site, te kiezen (zoals: "X is een soort Y"). De betrouwbaarheid van de omschrijvingen wordt gecontroleerd door te kijken naar het aantal correct gegeven omschrijvingen, de woordenschat en taalgebruik van de kinderen.

In drie weken tijd hebben 123 kinderen het spel gespeeld. Tientallen kinderen hebben het spel meer dan eens gespeeld, dus de kinderen lijken het spelletje leuk te vinden. De woordenschat van kinderen van tien tot twaalf jaar is goed genoeg om de gegeven woorden te omschrijven. Ook bevinden de meeste van de deelnemers zich in de "concreet-operationele fase" van Piaget, wat inhoudt dat kinderen op deze leeftijd zich meer bezighouden met hier-en-nu situaties en concrete feiten. Dit kan voordelig zijn bij het omschrijven van woorden. Ook blijkt dat kinderen op de leeftijd van 10 inderdaad meer concrete woorden omschrijven dan de oudere kinderen. Over het algemeen waren de omschrijvingen van de kinderen betrouwbaar en het aantal spellingsfouten en typfouten bleken laag. De moeilijkheidsgraad en de concreetheid van de woorden maakten nauwelijks verschil in de juistheid van de gegeven omschrijvingen. Kinderen kunnen dit spel dus spelen met verschillende uiteenlopende woorden die in het corpus staan en zij zullen deze woorden over het algemeen op een betrouwbare manier omschrijven. Als dit spel inderdaad zou worden gebruikt om het OMCS-Dutch corpus te helpen opbouwen, dan kan veel betrouwbare kennis in een relatief korte tijd worden toegevoegd.

Appendix A

Experiment data

The data gathered during this experiment can be downloaded from:

`http://ilk.uvt.nl/gezondverstand/GezondVerstandData/`