

A Kids' Open Mind Common Sense

Solving problems in commonsense computing
with a little help from children.

Pim Nauts

Bachelor Thesis, CIS (BDM)

Supervised by

Prof. Dr. A.P.J. van den Bosch

Tilburg University
Faculty of Humanities
Tilburg Initiative for Creative Computing
Tilburg, The Netherlands
December 2009.

Oliveoil is made of olives

Sunfloweroil is made of sunflowers

- then what exactly is babyoil made of?

Abstract *In this research, we combine a large collaborative approach to the problem of commonsense reasoning, a new approach to giving computers human-like reasoning in the field of artificial intelligence. By using children as contributors, motivating them with game elements and using in-game human validation in order to evaluate whether or not children can be valuable and reliable partners in building commonsense databases, we designed a web-based experiment with 10 to 12 year olds, showing children can contribute reliable commonsense to a commonsense database for Dutch. In a second experiment, conducted in classrooms, we evaluated a way of automatically validating the data children provide us with. We conclude children are a reliable source in collecting commonsense, automated validation can validate at least 30% of the data provided without human interference and more of the effort in validating can be channeled by using less assertions for guessing.*

Tags Artificial intelligence, developmental psychology, commonsense reasoning, semantic networks, Games With a Purpose, automated validation, Open Mind Common Sense.

Contents

Preface	6
1. Introduction	7
1.1 Computers are plain stupid - but that is just commonsense	7
1.2 The nature of commonsense	8
1.3 Digging for common sense	9
1.4 What do we know?	10
1.5 Turning to children	10
1.6 Approach	11
1.7 Research Question	12
2. Research on commonsense reasoning	13
2.1 History in commonsense research	13
2.2 Open Mind Commonsense	16
2.3 Where OMCS stands out	18
2.4 Commonsense development in children	18
2.5 Jean Piaget	19
2.5.1 Stages in a child's development	20
2.6 Infantile amnesia	20
3. Commonsense: Artificial intelligence needs natural intelligence	22
3.1 GWAP: Games with a purpose	22
3.2 Games as a framework for collecting data	22
3.3 In-game validation as a solution to truthfulness	23
3.4 Commonsense development within our subjects	23
3.5 ILK's Gezond Verstand	23

4. Methods	24
4.1 Experiment 1	24
4.1.1 Methods	24
4.1.2 Results	28
4.2 Experiment 2	29
4.2.1 Methods	29
4.2.2 Results	34
5. Conclusions	38
5.1 Limitations	39
5.2 Future Research	40
5.3 Final words	41
7. References	43
8. Appendices	46
8.1 Screenshots	46
8.2 Questionnaires	52
8.2.1 Explanation	52
8.2.2 Example stimuli (clue3)	53

Preface

I remember quite well how starting my studies a while back was a big relief to me. However hard it can be at times, at least this is my decision. And I've been very lucky to choose Tilburg and TiCC. Lucky to choose for such an inspirational environment, being rewarded for curiosity.

Writing this thesis itself fell quite hard. It has been a lengthy process from first initial conversations over the course of the experiments to finishing the piece of writing; a process in which I came across many mirrors. I got pulled into an exciting and challenging project and I had the luck to write my thesis on such a fascinating subject. Thank you Antal, for showing me your world and views and your never ending patience and enthusiasm. Thank you Nienke, for all the fun and laughter and being one of the guys. Thank you Peter, for coding our wildest dreams and all the hours you put in. Mama, Papa - you know what.

I want to express my gratitude to those who made it possible to conduct our experiments: Marga van Zundert from the Kinderuniversiteit, Ria from BS De Regenboog for her fantastic interference and help. Juf Pascalla & Juf Georgette van OBS de Twijn for their cooperation even on such short notice. Hendrik, you've earned lifelong coffee.

Aaf, thank you for keeping me on track and all else there is.

I've been able to do our research in an environment that is as inspirational as accessible. I did my very best to reflect just that in this thesis and I hope this is a comprehensible introduction into what might turn out to be part of our future.

1. Introduction

Let's think, said Amelia Bedelia. *What can we do together?*

As they talked, a truck pulled into the driveway.

Excuse me, said the driver. Is this eighty-one Fairview Avenue?

No, said Amelia. This is eighteen!

Oh, said the driver. I thought I recognized this house from the plans.

What plans? said Amelia. And who are you?

My name is Bill, said the driver. And this is Eddie. We are in construction, the driver said

In construction?! said Amelia Bedelia. You look like you are in a truck!

- Herman Parish - Amelia Bedelia

Amelia Bedelia is a French maid who takes language a bit too literal. She ploughs her way through a hardly understandable field of reference, inference and ambiguity. Though a bit silly sometimes, you cannot fail to notice that she is right when she is wrong. How can we blame her if no one ever told her about the facts she apparently misses out on?

A truck is a type of car / Trucks can park on driveways (and drive on parkways) / there is a difference between being in construction and being under construction.

Amelia Bedelia's view on the world in a way resembles computers' view on the world. Computers cannot deal with ambiguous input and no one ever told them how to deal with it. Just like Amelia Bedelia.

In itself, that is a strange thing. Computers are powerful machines that have a very important place in modern life, outperforming humans in many ways. However, computers are still lacking in the everyday situations and interactions which are the easiest of tasks to humans - try to get your computer to tell bananas from tomatoes and then ask a random 4-year old.

1.1 Computers are plain stupid - but that is just commonsense

In many ways, we would like our computers to behave as digital assistants to everyday life, we want them to autonomously perform the type of task or assignment even little children can understand. Would it not be great if Google actually answered a query instead of referring, or when your cellphone automatically scheduled the meeting you just made an appointment for (Mueller, 2000)? Why are humans the ones that need to adapt in human-computer interaction. Despite years of research into computer science and AI, building machines that can think about ordinary things the way humans do are still out of reach. Why is it we seem unable to make computers think about the world? (Singh, 2002).

Computers do not know about the world from a functional, procedural and relational perspective. Humans, on the other hand, do. We have invented words to refer to objects and concepts, how they behave and how they relate to each other, we know how to solve apparent ambiguity by

deriving proper meaning from context. Clearly, computers cannot. What 'we' got and 'they' lack, is knowledge about the most ordinary things in the world. Things so ordinary we tend to forget about our understanding and do not even realize we know. Were we able to give computers such knowledge, were we able to give them ways to represent how we think and reason about the world - then they can be made to become personal assistants and function autonomously in the human world (Singh, 2002).

1.2 The nature of commonsense

Marvin Minsky explains the nature of the type of knowledge concerning acquiring complex skills in one of his best-known books, *The Society of Mind* (1988):

"The mental skills we call 'expertise' often engage large amounts of knowledge but usually employ only a few types of representations." (p. 327)

So, in other words: performing a complex task requires deep understanding of only a narrow domain. There is a lot to know and you need to be skilled to play a game of chess. But it is an exhaustive, relatively compact area and skills needed are very specific; given very robust algorithms we can mimic a game of chess in lines of code. Commonsense reasoning, on the other hand, requires knowledge about a very broad domain and the types of presentations involved, applied in a very implicit way.

Minsky *"Commonsense thinking is actually more complex than many of the intellectual accomplishments that attract more attention and respect (...) In contrast, commonsense involves many kinds of representations and thus requires a larger range of different skills."* (*The Society of Mind*, p. 327)

What we, as humans, consider easy is in fact not so much a deep understanding of a narrow field, but much more the shallow understanding of a very broad field. It is those aspects of human behavior we most take for granted that are hardest to emulate on a computer (Mueller, 2006). The problem with giving computers commonsense therefore is not about figuring how to make a particular method of reasoning work over a particular type of knowledge, but how to make systems that can cope with many types of knowledge and many types of reasoning over them (Singh, 2002). Since there is so much to know about the world, so much commonsense, key is in the volume of input needed to construct such reasoning.

Commonsense, in other words, is knowing little about much. As far as today, computers have learned much about little.

Intermezzo: commonsense put into perspective.

What exactly is regarded common sense?

Wikipedia (...) that which [humans] sense (in common) as their common natural understanding.

Einstein: (...) the collection of prejudices acquired by age eighteen.

Marvin Minsky: *"The mental skills that most people share."* (p. 327)

Commonsense, ranging from knowledge about objects and relationships to procedures, is an immensely complex construct in our minds for which we need a large part of the first half of our lives to acquire them.

In (adult) conversation, commonsense is all not explicitly mentioned anymore. However, we also use commonsense in solitary acts (thinking, mesmerizing, acting). Our functioning is based on an unexplainable but obviously apparent base of knowledge. However vague, we do understand the world surrounding us. Everything we say, hear, write or read has that underlying layer of understanding, 'extra' knowledge we build upon and derive meaning from that seems to be forgotten but at the same time does enable us to understand. It is so common we commonly forget about it.

If you think about it, in an everyday situation, e.g. if I order a diet coke, what is implicitly referred to that does not even surface our thoughts?

- I'm probably in a restaurant or bar (what's the difference between a restaurant and a bar?)
- I'm thirsty (what is thirst? What does one do when thirsty? How do I know I'm thirsty?)
- I'm talking to a waitress (what's a waitress, how do I recognize one?)

I recently was at an airport, waiting in line at a fast-food restaurant to order a milkshake. In front of me, an American ordered a large Fanta (soda) and was genuinely astonished by how small 'large' is here in Europe. Sometimes, commonsense can be sensed uncommonly.

And then there is the problem of how we know to distinguish between similarities or the need to know when the obvious is not quite so obvious. How can one make sense without extra knowledge when we fill in forms by filling them out, find that quicksand takes you down slowly and boxing rings are square. And, if a vegetarian eats vegetables, then what does a humanitarian eat!?! (taken from Gray, 2003).

1.3 Digging for common sense

People are constantly learning - about the world and about how to apply this newly acquired knowledge to the world. For instance, fact X plus fact Y gives action Q. Imagine it is raining (X) and I need to get go to university by foot (Y). Then I will probably bring an umbrella (Q).

While my inference sounds quite logical and natural given the premises, it is a completely autonomous process in which I build my reasoning on known but seemingly forgotten facts (the problem that arises from rain + going outside can be solved by bringing an umbrella since umbrellas are ought to protect you from rain).

When we want to teach our computers the commonsense we all know about, the first step is to explicate that commonsense. And that is a problem: commonsense is taken for granted since we often do not recognize how intricate those facts and processes are (Minsky, 2006). In other words: commonsense is hidden somewhere deep in our minds and it is really hard to dig for.

On top of that, commonsense is a domain of great scale that has been regarded as simply too big to tackle with respect to giving computers commonsense (Singh, 2002). Several attempts have been made at estimating exactly how much commonsense people have, ranging from lexicon estimates extended by number of known facts to neuron counts multiplied by the number of bits a neuron can store (Singh, 2002; Landauer, 1986; Von Neumann, 1958; Minsky, 2006). Though the outcomes varied, we can confidently state there are many millions of commonsense facts apparent in our minds.

1.4 What do we know?

How much information can we actually store in our brain? Marvin Minsky (2006) suggests that when we account for our psycholinguistic system, object knowledge and social realm, a "humanlike reasoning system" requires hundred millions of items of knowledge. Thomas Landauer chose a quantitative and more precise approach and proposed the idea of using the smallest piece of information possible; a "bit". He found that human commonsense knowledge is probably around 1,000,000,000 (one billion) bits in size (Landauer, 1986). Either way, it is a problem of great scale and magnitude.

Will we be able to teach our computers commonsense, we will need to find a source that contains all this commonsense. While commonsense is an apparent part of our everyday thinking, it is not an apparent layer in our everyday thinking. It is not very likely - nor true - that we have written down our commonsense somewhere: commonsense is nowhere to be found in a recognizable form. Despite the accuracy and potential of a collaborative approach like Wikipedia there is nothing, not even on the web, that can be used to retrieve the very basic commonsense we are looking for here.

But still, we do know these many millions of facts, which we acquired at some point during our lives. I had good reasons in using the analogy with the 4-year old I used before: to gather the commonsense facts we are looking for, maybe we should look at those right in the middle of the process of acquiring them - we should turn to children to teach our computers about the world.

1.5 Turning to children

During childhood, we have acquired most of the commonsense used throughout our lives. But that is a tedious and long process of trial and error, and maybe the most implicit of all our learnings.

To children, commonsense is an essential part of their everyday reasoning and they are explicitly expanding their mental databases everyday. We should use this natural curiosity for building a commonsense database and ask children to teach our computers about the world.

Turning to children to leap around the implicitness of commonsense leaves us three main problems:

- Children are very naive in their views on the world. Thus, to a certain extent, we may not assume they will provide us with truthful commonsense and reliable assertions. In Chapter 4

(Commonsense development in children) we elaborate further on that matter. Therefore, we must think of ways to account for a lack of truthfulness in their assertions.

- Children are right in the middle of learning to express themselves. Most adults find writing and correct spelling a hard thing to do, let alone children. Therefore, we must account for misspelling and misunderstandings as well.
- There must be a way of getting children to contribute to our problem of commonsense in an implicit way, children are not a resource of labour; it is not likely we will succeed by offering them candybars in exchange for their help (apart from ethical concerns, even).
- Commonsense is a problem of great scale.

Are there ways to overcome these problems?

The problem of truthfulness is of a vital importance: would we already be automatically able to determine the truthfulness of assertions given a certain word or concept, we would not need to conduct this research. We propose another way to automatically rate - accept or reject - the assertions children have given us. The current state-of-art in natural language processing does provide us with tools to automatically compensate for misspellings and lexical problems, on which we will elaborate in further sections. By adding game elements, we can motivate contributing children in helping us explicate commonsense.

That leaves us the problem of scale. As we will see in further sections, there has been a recent trend in AI-research to use collaborative approaches in solving large-scale problems at which humans still outperform computers (Stork, 1999), e.g. in adding game elements (Von Ahn, 2007; Von Ahn et al., 2006) by combining computational power and natural intelligence.

1.6 Approach

In this research, we have combined a large collaborative approach to the problem of commonsense reasoning with the use of children as contributors, motivating them by the use of game elements. We apply in-game human validation as a means to evaluate their contributions and whether or not children can be valuable and reliable partners in building commonsense databases. Not much research has been done before on this topic, therefore we are mainly concerned with exploring the possibilities and want to know if our approach proves to be a potentially promising one.

We designed a two-phase game aimed at collecting common sense facts, input will be collected in a database. In the game, children will be given a concept to describe (e.g. "boat"), for which they can choose from a number of templates (e.g. "isA"). In the second phase, children are asked to guess the given concept based on another user's description ("e.g. ... isA mean of transport.") (please proceed to experimental design for a thorough overview of the game setup). In solving the problem of truthfulness (the process towards a valid, sensical and truthful piece of commonsense knowledge), we incorporate this as a validation-component to let the children, without knowing they do so, asses and judge each others' assertions, resulting in confirmed truthfulness.

Using the assertions children have given us in the first phase of the research and using the results for the second phase (reliability-testing) we want to answer our research question(s).

1.7 Research Question

The main research question, its subquestions and a subsequent hypothesis are as follows:

Main Question *Can children (ages 10-12) be a reliable source in adding commonsense facts to a commonsense database, can their contributions be automatically validated and can we raise the efficiency of their efforts?*

Subquestion *Can in-game validation be a means for automated acquiring of truthful commonsense using children?*

Subquestion *Does the number of assertions used for guessing the target concept influence reliability?*

H1: The number of assertions used for guessing affects the correctness of the answers provided.

2. Research on commonsense reasoning

Within the field of AI, commonsense has been recognized as a bottleneck problem since the early 1960s (Kuipers, 2004). The idea that computers' inability to cope with human behavior is due to a lack of commonsense was first proposed by John McCarthy in a classic 1959 paper. McCarthy and Marvin Minsky, among others, have since developed an extensive theory on the need for commonsense in human-computer interaction from both formal and human perspectives (McCarthy, 1959; McCarthy & Hayes, 1969; Minsky, 1986). The idea of the need to build upon commonsense has led to a number of attempts to design a suitable architecture enabling a useable database and building concrete applications.

2.1 History in commonsense research

Though not an exhaustive overview, what research has been done in the field of commonsense reasoning?

Mindpixel Digital Mind Modeling Project MindPixel was launched by Chris McKinstry in 2000. MindPixel was built upon a database and algorithm known as GAC (Generic Artificial Consciousness). Every websurfer could add assertions, which McKinstry dubbed MindPixels, to GAC through the MindPixel-website. In contrast to the original OMCS, McKinstry used human validation to rate the MindPixels on truthfulness (true-sometimes.....rarely-false), based on his proposal for a revision of the Turing Test, the Minimal Intelligence Signal Test (MIST) (McKinstry, 1997). In 2004, this approach produced a large database with 1.4 million MindPixels. McKinstry ended the MindPixel project in 2005, an ample year before he committed suicide. At the time of writing, hosted by an internet portal called I am bored (sic!), MindPixel lives on by the face of an 'IQ-test'. One can join in and rate assertions taken from GAC by a probabilistic measure ('does invade produce gas' - doubtful).

ThoughtTreasure Erik Mueller, an MIT Media Lab-researcher, initiated ThoughtTreasure in 1993, a "comprehensive platform for natural language processing and commonsense reasoning" (Mueller, 2003). Launched to the public in 1996, "51,305" assertions were collected within the lifetime of the project.

In ThoughtTreasure, concepts are organized into a hierarchy, e.g. evian is a type of flat-water, which is a type of drinking-water, which is a type of beverage, which is a type of food, etc (Mueller, 2003). Mueller was inspired by Doug Lenat's CyC (see below) to the extent of using a formal language to represent knowledge to the database (avoiding ambiguity and simplifying interrelating concepts). However, where Lenat uses mostly logics ThoughtTreasure also uses scripts, "representations of typical activities" (Mueller, 2003).

In 1997, Mueller founded SigniForm, aimed at exploiting the commercial possibilities of applying narrower, domain-specific commonsense reasoning in software, e.g. SensiCal, an application that can apply commonsense to digital calendars - "You are taking Lin, who is a vegetarian, to a steakhouse." (Mueller, 2000). SigniForm's activities were brought to a halt in 2000 when it did not turn out to be a viable venture. ThoughtTreasure has not been developed further since.

Cyc Probably the largest and best known of projects in the field of commonsense and AI is Cyc, started in 1984 by Doug Lenat. In 1994, Cyc stepped into the real world when Lenat founded Cycorp, a company aimed at commercially exploiting Cyc's potential. In the 10 years of development Lenat developed an architecture and commonsense database that consisted solely of manually entered knowledge. This was done by what he dubbed "knowledge engineers" using CycL, a formal language specifically designed for Cyc. Cyc could then connect (the concepts within) the assertions entered and create inferences.

In about 2 decades, Cyc gathered over a million facts. Though this is a remarkable effort in itself, it also points out Cyc's main weakness: contributions can only be made by experts, which limits 1) the number of contributions that can be made and 2) the time it'll take to cover the whole domain of human commonsense. Next to that, developing Cyc has been a very expensive undertaking (50 million dollars plus) that did not yet yield satisfactory results given its funding. However, at the time of writing, Cycorp has released Cyc to the general public (under a Creative Commons License) as OpenCyc and has started incorporating Cyc-contents and technology into semantic web applications.

That Cyc in its initial form still exists today and has proven to be a promising approach to commonsense reasoning is a remarkable and widely acclaimed fact in itself. Though Cyc's approach can specifically be applied to narrower, domain-specific commonsense reasoning, we still believe Cyc is dependent on specialists to cover the many millions (hence Landauer) of commonsense facts.

Open Mind Initiative A radical different approach, not so much aimed at solving all the problems that arise in representing human knowledge to computers (thus accepting that is also part of human commonsense), is Open Mind commonsense (sprung from the Open Mind Initiative). With the promise of much easier harvesting much more commonsense facts, this is what we consider the most fertile approach to commonsense within AI.

The Open Mind Initiative was initiated within the MIT Media Lab by David Stork in 1999. Open Mind is a large-scale web-based framework for collecting data contributed by non-experts through collaboration (Stork, 2004), aimed at creating intelligent software.

What sets OMI apart from the 'usual' approach, e.g. from Cyc, is the use of non-expert contributors - OMI adopts a bottom-up approach and uses the growing number of nonspecialist users ("netizens") found online, with the Web as a low-cost framework for collecting data contributed by large numbers of users (Stork, 2000). These data can then serve as a base for intelligent software, e.g. in the field of speech recognition (Valin and Stork, 1999).

The largest project within OMI is Open Mind commonsense. Initiated by Marvin Minsky and Push Singh in 1999, it applies the basic idea of exploiting the web's userbase to collecting commonsense facts through the web. OMCS proved to be a very interesting approach to the problem of gathering commonsense facts, with around 700,000 English facts in the database from 14,000 contributors between its launch in 2000 and 2004, at only marginal costs compared to Cyc (money goes into the development and design of the architecture, not into entering data).

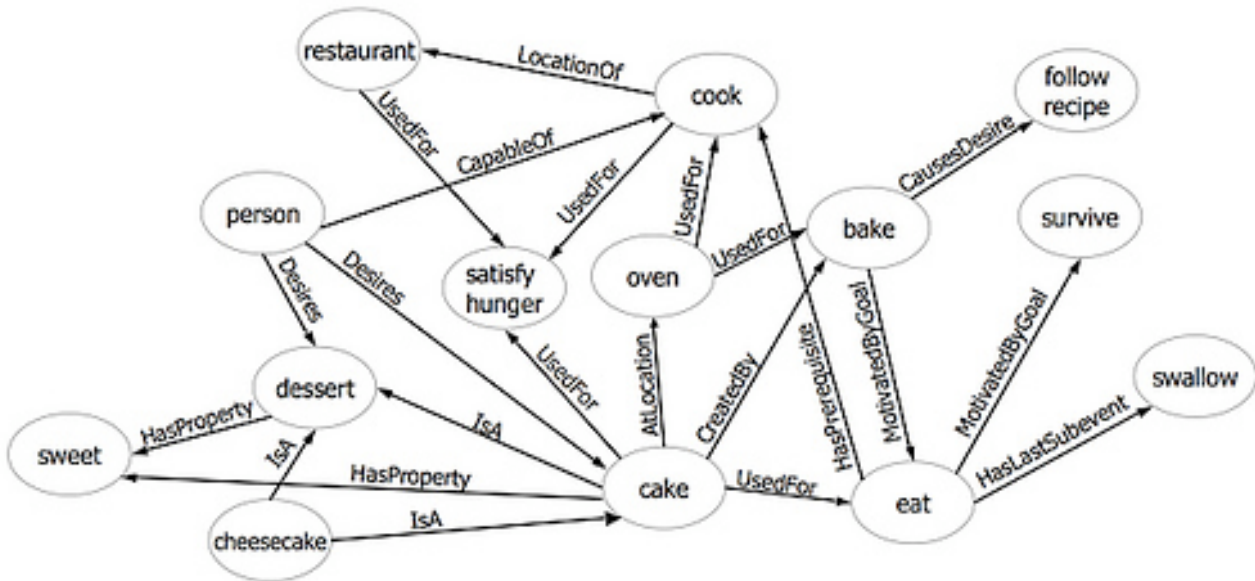


Figure 1 Conceptnet’s architecture

The data OMCS gathers are added to another OMI-project, Conceptnet, a simple, easy-to-use semantic network, in which concepts are related through predicate types (object-relation in a proposition). These predicates are extracted from the templates OMCS users have filled out and reworked into computable data.

Semantic networks consist of nodes which represent concepts and links which represent predicates. Meaning comes from the interconnected nodes and the type of connection between them: a boat is made from wood is made from trees.

Another semantic network, like Conceptnet, is WordNet; a lexical database of English in a semantic network that defines the meaning of words in a hierarchy (holonymy, meronymy, etc) using word sense - the flower used in pasta is not the same flower as the type of thing that cheers most women. It differs from Conceptnet in the way that it does not use semantic relations between concepts. In Wordnet¹, concepts are part of a larger concept or converging - sunflower is a flower is a plant is a vegetation is a lifeform.

¹ For more technical details on Wordnet it is Wikipedia-lemma provides an excellent resource at <http://en.wikipedia.org/wiki/WordNet>.

2.2 Open Mind Commonsense

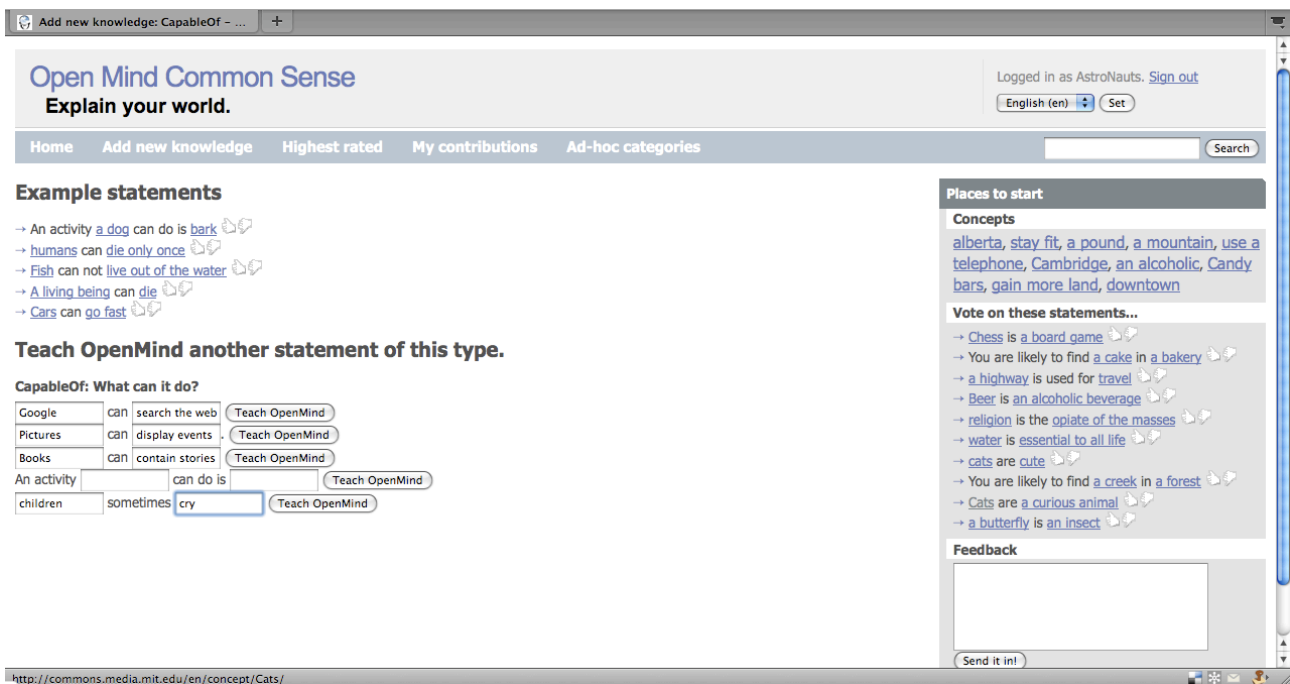


Figure 2 With OMCS ordinary 'netizens' can add commonsense facts

As OMCS aims at harvesting the web's potential and scale, they adopted a semantic free-link structure, as opposed to the common rule-based approach in AI (Singh, 2002). For it to be usable to non-specialist users, you would want the possibilities for adding knowledge to be as little different from what people are used to as possible - you would want them to use their own language.

However, using natural language as input does give a few problems. At first, commonsense is a very broad domain and the types of knowledge vary to a great extent. You would therefore need your system to be able to cope with all these different types. Secondly, main problem with natural language is that computers require precisely and accurately expressed chunks of knowledge whereas vagueness and ambiguity lie in the very nature of human language. Therefore, "if ambiguity cannot be avoided, then we must learn to cope with it" (Singh, 2002). Rather than directly engineering knowledge structures used, commonsense is extracted from the natural language-input contributors provided (Singh et al, 2002). The state-of-art in natural language processing was, at the time OMCS took off, regarded robust enough to be able to extract the commonsense from users' non-formal input. Since 1999, two versions of OMCS have been online: OMCS-1 and its evolution, OMCS-2

OMCS-1 In the very first version of OMCS, users could add knowledge in ordinary English sentences by stating facts (coffee keeps you awake), filling in explicit templates to describe a certain concept (the effect of coffee is), presenting short narratives and asking to describe what can be derived from it (Flinn was drowsy. He poured in some coffee.) and presenting photos and describing the events displayed (President-elect Obama holds a Starbucks-cup | Obama is having coffee).

Since OMCS did not require contributors to use a logic template to add knowledge, commonsense had to be separately extracted from the input. In a 2002 paper, Singh et al describe how they did a first set of experiments to evaluate the contents of the OMCS-1 database by using part-of-speech tagging to divide input in syntactic components and form them into standardized representations.

OMCS-2 With the data in OMCS-1 and its evaluation, the team began working on the next-gen OMCS, which mostly tackled the deficiencies in OMCS-1 (Singh et al, 2002).

In OMCS-1, the team found contributors vary in the way they like to enter knowledge. "Some [contributors] like to enter new items. Others like to evaluate items. Others like to refine items". In OMCS-2 the process towards a single piece of commonsense knowledge is therefore divided into four different stages, all of which each user for each contribution could perform, aimed at validating the input and inferences.

The predicate templates users can choose from in OMCS-2, range from simple descriptives (what is it made of?) to obviously harder procedural and declarative templates (how do you bring it into existence?). Based on these templates, users sort-of answer a question. These assertions are parsed into meaningful parts (concepts) and interrelated. The system can then relate these concepts and produce assertions. The input entered into the templates can consist of either noun phrases, adjective phrases, verb phrases or prepositional phrases. A small selection of available predicate templates in OMCS-2 can be found in [table 1.]

Relation-type	Question posed	Syntactic form
MadeOf	What is it made of?	NP is made of NP
IsA	What kind of thing is it?	NP is a kind of NP
UsedFor	What do you use it for?	NP is used for VP
CapableOf	What can it do?	NP can VP
PartOf	What is it part of?	NP is part of NP
CreatedBy	How do you bring it into existence?	You make NP by VP
HasPrerequisite	What do you need to do first?	NP requires VP
AtLocation	Where would you find it?	NP is found at NP

Table 1 Predicate types used in OMCS

OMCS in different languages OMCS applies a basic idea to the English language, eventually collecting all commonsense in, yes, English. However, this basic idea goes for each and every language spoken in the world. Off-course, the number of native speakers makes for English to be a logical choice (when there are a lot of native speakers, the few willing to contribute still come in much large numbers than the few you'll find for less widely spoken languages). Therefore, for future commonsense applications to be usable in other languages than English, a language-specific commonsense database is needed. If these exist, it is not hard to imagine it can also serve as a framework for, e.g., cross-cultural communication addressing the many cultural differences faced in

cross-cultural communication (Chung, 2006). Currently, there are OMCS's in Portuguese, Korean, Japanese and Dutch.

OMCS 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, contents of which can be used to initiate commonsense projects in Dutch. Since OMCS Dutch is new and has not received very much attention yet, its contents are extremely limited.

While OMCS is one of the most promising commonsense projects to date, it does not deal with what is regarded as one of the toughest technical problems in commonsense reasoning: organizing and contextualizing assertions (Singh and Barry, 2003) to avoid ambiguity - a problem Doug Lenat tackled by putting assertions within a well-defined context capturing underlying assumptions (Singh and Barry, 2003). However, within OMCS it is believed those problems can be solved by looking at the contributors' context (Liu and Singh, 2004) and its key difference lies in the adoption of diversity in human commonsense and the use of natural language as well (Singh, 2002).

2.3 Where OMCS stands out

In OMCS-2, users can still add free-form sentences next to the template-based input. Main point of interest in OMCS-2 to us is the addition of rating the truthfulness of an assertion both from other contributors as well as an inference the OMCS-system came up with. As stated earlier, the potential of using children as commonsense contributors lies in their views on the world, right in the middle of acquiring information about the world and thus more conscious of what's regarded obvious by adults. However, as goes with adults but to a larger extent, this is a process of trial and error and children are not infallible. Therefore, applying the validation built into OMCS-2 into the process and thus asking separate contributors about its validity before it is considered knowledge (rather than afterwards), would raise the quality of these pieces quite dramatically.

2.4 Commonsense development in children

Developmental psychology has known several paradigms on children's cognitive development in the last decades (Moore, 2006), ranging from the tabula rasa or blank slates to preprogrammed and innate skills present from birth on (nature/nurture). In *The Society of Mind* (1988) Marvin Minsky gives his views on cognitive psychology in children. According to Minsky, children both learn from experiences and have the innate abilities to give basic structure to these experiences which, over time, leads to logical abilities. This is what Minsky calls Papert's Principle (by the name of AI-hero Seymour Papert): acquiring knowledge is not simply accumulating facts; "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" (Minsky, 1988). In [Figure 3], this is illustrated by the heuristics applied: "The new Appearance administrator is designed to say "more" when the agent Tall is active, to say "less" when the agent Thin is active, and to say nothing at all when something appears both taller and thinner. Then the other new administrator, History, makes the decision on the basis of what Confined says."²

² Taken from <http://www.papert.org/articles/PapertsPrinciple.html>, retrieved July 28th 2009.

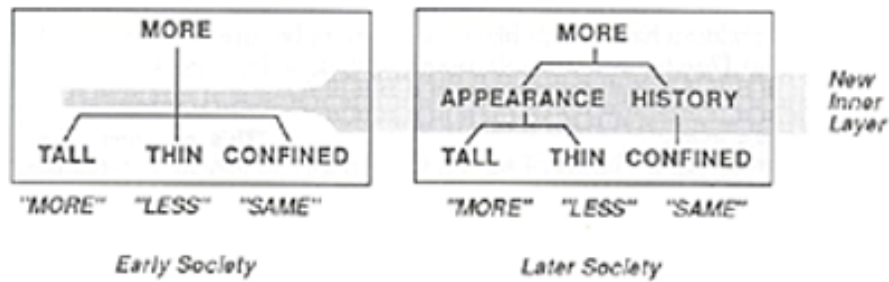


Figure 3 Papert's principle explained.

Seymour Papert was especially interested in how cognitive development in children could be supported by the help of computers and is a great advocate of educational computer-based technology (Eckhardt, 2008). Drawing upon that idea, Papert designed "Logo", a programming language with which children can practice and improve their skills in problem-solving, significantly contributing to the test-scores of the children involved (Papert, 1980; 1944; Harel and Papert, 1990). In the seventies and eighties of the twentieth century, Papert built small robots using LEGO (the Danish producer of colorful toy-bricks), on which he published the book *Mindstorms: Children, Computers and Powerful Ideas* (1980). LEGO used this for their own series of Mindstorms robots later on^{3 4}. Papert's ideas about cognitive development in children closely draw upon the ideas proposed by child psychologist Jean Piaget.

Jean Piaget is a well-known name in psychology and largely pioneered the field of cognitive development in children. Piaget carried out many experiments with children and his results gave a base for understanding children's cognitive development (Eckhardt, 2008). Piaget focused on distinctive reference at certain ages, e.g. the ability to make correct estimates and assumed that knowledge is acquired through interaction with the surrounding world.

2.5 Jean Piaget

According to Piaget, there are two principles that extend intellectual capacity and guide growth and development of a child: organization and adaptation. Organization refers to innate capacities to structure incoming information with the help of schemata; feature sets and characteristics of what a certain concept or object is associated with. According to Piaget, a child's world view is organized in these complex and integrated ways (Bhattacharya and Kahn, 2008) and schemata are used to understand and structure the world

With what he refers to as adaptation, Piaget believes we use mental structures on which we try to fit external events. These mental structures are a variable that constantly changes and adapts to a changing environment and surroundings. Adaptation consists of assimilation and accommodation.

Assimilation Assimilation is the fitting of existing internal schemata or operations to the external reality, e.g. when a child identifies each animal as a 'dog'. Piaget believed that assimilation, from a functional perspective, is very important to a child's development - it enables a child to grasp

³ http://en.wikipedia.org/wiki/LEGO_Mindstorms

⁴ A worthwhile read on Mindstorms, from *Wired*, is [The Origins of Mindstorms](#) by Jim Bumgardner.

reality based on what's already part of that reality, thus there is only the need to adapt (assimilate) an existing schema instead of creating and learning a complete new one.

Accomodation With accommodation, Piaget tries to explain what he believes is the process of changing internal mental structures to match reality when existing schemata or knowledge cannot account for a certain occurrence. Literally: accommodating to newly presented situations.

2.5.1 Stages in a child's development

Within his body of research, Piaget identified four main stages in children's cognitive development; growing from very different worldviews to the same mental routines adults have (Atherthon, 2008).

Sensori-motor stage *Birth to two years old*

Sense and motor skills learn children to understand the world through trial and error ; knowledge is based on physical interactions and experiences, therefore limited.

Pre-operational stage *Two to seven years old*

Use of language, memory develops. Links between past and future are identified and expressed; complex relationships as cause and effect are not yet grasped. The child is the centre of its world, intelligence is egocentric and intuitive, not logical (Wood et al, 2001). Imagination plays a very important role in solving problems in current reality .

Concrete-operational stage *Seven to eleven years old*

Logical inference develops, but only related to concrete objects and events.

Formal-operational stage *11 years of age and older*

From adolescence through adulthood humans develop the use of symbols related to abstract concepts and can think about multiple variables in systematic ways.

2.6 Infantile amnesia

According to Piaget, every child goes through these four stages in subsequent order, with ages slightly differing. During these processes, the commonsense corpus also extends and develops. From basic sense- or motor-derived facts (hot/cold, heavy/light) trough causality and logical inference. In terms of Piaget's developmental stages, we could identify a certain commonsense fact and a logical inference according to these stages. Though an obvious result of developing ourselves as young birds, these facts are so highly internalized we are not aware of them as grown ups.

Fact *The beach is near the sea*

Stage	Activity	Learnings
SensiMotor	experience sand and water, taste salt	element characteristics
Preoperational	element reference, naming; play	verbal concepts, interaction
ConcreteOperational	play, combinations (wet sand)	logical inference; causality
FormalOperation	understanding (tides, liquid sand	all combined

In real life, we do not think of a beach as such. Nonetheless, it is part of our commonsense database, suddenly surfacing when we see our kids play on the beach just the way we did.

Many of the abilities adults have developed during their early childhood. But why is it we can easily remember events later on in life but are unable to recall events from early childhood (Eckhardt, 2008)?

In *The Emotion Machine* Marvin Minsky (2006) calls this infantile amnesia. Minsky argues that while we are able to internalize the facts concerning certain events we lack skills to remember and reconstruct these memories. Yet, in these earliest years (Sensory-Motor and partly PreOperational stages in Piaget's terms) we acquire most of our basic skills and basic commonsense (Eckhardt, 2008). As a result, at ages we are able to reconstruct events and memories, we cannot explain how we memorized them.

That poses a very relevant problem for the harvest of commonsense: those who know, do not know they know and are not able to retrieve and recall. Therefore, we turn to children.

3. Commonsense: Artificial intelligence needs natural intelligence

A problem commonly faced in computer science is that traditional large-scale problems, such as commonsense, require the joint capabilities of both computers and humans: humans individually cannot provide the vast amounts of information needed whilst computers have no clue about the information itself but are able to process. Traditionally, the focus in solving such problems has been on improving AI-algorithms, starting from a top-down view on how problems can be solved. In our approach, we have adopted a different idea by combining the areas in which both computers and humans have unique capabilities, being the computational power of the first and natural intelligence of the latter.

With OMCS, a viable effort to solve the problem of scale has already been made. Nonetheless, question arises whether the approach of OMCS is the most effective: it is heavily depended on the effort of willing contributors, donating (much) of their time to a project that can only yield results from a very large scale of input. Would we want to solve the problem of commonsense regarding scale, we should appeal to a much larger public than just those interested in contributing to AI. Furthermore, the effort made by contributors to OMCS is not as efficiently used as could, specifically because contributors themselves have many of choices in what concepts/types of knowledge to contribute and because validation -we would want useful, truthful facts, is part of the website, not an integral part of the process of contributing. Would it be, every contributor would yield useful, compact knowledge for ConceptNet.

In sum, we should address much more people and use the effort made more efficiently.

3.1 GWAP: Games with a purpose

Our solution to the non-technical problems in commonsense computing is derived from an idea proposed by Luis van Ahn⁵. and Laura Dabbish, researchers in Carnegie Mellon's CS department. They proposed a novel approach to solving these type of problems: GWAP's | Games With A Purpose. In using the constructive channeling of human brainpower through computer games, data generated as a side effect of game play also solve these computational problems and helps improve AI algorithms (Von Ahn and Dabbish, 2008). This way, the framework of collecting data appeals to a much larger audience, regardless of their interest in contributing to AI (a setup that is intrinsically much more motivating then the endless adding and rating of assertion in efforts like MindPixel or OMCS as discussed).

3.2 Games as a framework for collecting data

As a first step in using human intelligence to solve large-scale computational problems, von Ahn developed the ESP game, an image-guessing game. Von Ahn faced the same problem we are

⁵ Luis von Ahn also developed the CAPTCHA (<http://www.captcha.net>) which is basically a 'reverse Turing Test': a way to tell computers and humans apart, broadly applied over the internet as a security measure to keep spambots out of (social) networks (Von Ahn et al., 2004)

dealing with here: truthfulness. The ESP game is designed to let people describe pictures, so to have a reference for algorithms in image recognition. However, the robustness of this algorithm is completely dependent on the quality of the player's input. How would von Ahn know if a player's input would really be objects recognizable on the images?

3.3 In-game validation as a solution to truthfulness

Very cleverly, von Ahn decided to pair his players with another, anonymous and random, player. Both carry out the very same task: describing the image they see. Players cannot see one another's input and do not receive points until they use the same concept to describe the very image. For every match, extra points are given. This way, if players independently agree on a certain object being present in the image, it must really be visible (von Ahn and Dabbish, 2004).

3.4 Commonsense development within our subjects

Aged ten to twelve years, most of the basic commonsense is already acquired and developmental 'level' of the subjects in our experiment will be somewhere between the concrete-operational stage and the formal-operational stage. Therefore, it can be expected certain problems as to reliability can already be excluded: egocentric descriptions (cars ought to be blue to be cars) or out-of-this world explanations belong to a past phase in development.

A final important aspect is the intrinsic motivation we hope our game will elicit: in the concrete-operational stage children are very keen on understanding the world surrounding them. We therefore expect them to be keen on playing the game as well.

3.5 ILK's Gezond Verstand

By combining both these approaches we can use data as a side effect of gameplay from a broad and large public and validate these data all in one try.

Add to that the innate inability of most adults to effectively explicate their commonsense knowledge as highlighted before and one could assume there are highways to the goals we want to realize: usable commonsense applications that can actually act as assistants in everyday human life.

Therefore we propose a commonsense game with a validation component, specifically aimed at children. By asking children to learn our computers about the world, we can overcome the internalization problem present in adults. By using a game as a framework for collecting these data, we can channel the collective effort made towards our own aims, while realizing other, more attractive, goals for the children contributing; data basically being 'just' a side effect. Certain problems that might arise in the use of children (truthfulness, egocentric perspectives) can be avoided by carefully picking certain age groups. Furthermore, as a side effect of a game designed specifically for children, it does appeal and can be played to all ages above as well .

4. Methods

Drawing upon the foregoing body of research and earlier projects, we propose a new approach to the collection and validation of commonsense facts as side-effect of gameplay with the use of children.

For the first experiment, done with Nienke Eckhardt, we designed an online game that enables children to explicate, in part, the commonsense of the world surrounding them and explain it to our computer with the use of predicate types. Using an online game does give certain problems to the controllability of the research context (we cannot eliminate look-ups, helping parents or let the computer act as a gatekeeper that recognizes non-research group contributors) but we believe an online game gives us the opportunity to reach a much larger group of children, addressing the problem of scale in commonsense more appropriately (thus a more viable approach at the expense of some noise in the resulting data over fully controlled settings). The first experiment consisted of asking children to describe given concepts.

The second experiment was aimed at determining the accuracy of the proposed validation-element by marks of reliability using the data from the online game. It was carried out in a real-world setting, using simple questionnaires.

4.1 Experiment 1

For a thorough overview of the first experiment and its specific research questions and results, see the original paper that reported about the first experiment from Nienke Eckhardt (A Kid's Open Mind Common Sense, 2008).

4.1.1 Methods

Participants 123 children aged 10 to 12, 63 male and 60 female. Children were invited to play the game through their parents by sending them snail mail, explaining some background, overcoming the consent problem and explicitly pulling them to the website. Addresses were collected from Children University's mailing list, an initiative in which children are taught classes from real professors. No specific efforts to control the consistency of the group were made.

Materials We build an online game to collect the commonsense efforts of the children. In the game a static fictional character, Robot Rob, was used to act as a personified agent to both guide the children through the game and to give body to the concept of 'computer'. Such agents have proven to have positive effects on credibility, motivation and perception of experience (Van Mulken et al, 1998; Ball et al, 1997; Lester et al, 1997). Several human characteristics (a human name, use of natural language (not...I errrr AM rrrrrr A brrrrrt ROBOT)) were bestowed to improve effectiveness.

The interface was build using a text-based centered design with bright, attractive colors, and as few attention-intrusive visual elements as possible. Input could be contributed through standard HTML-based forms. Elements were analogous (in a playful way) to the robot/computer.

For screenshots, see appendices or see for yourself online at <http://www.ilk.uvt.nl/gezondverstand>



Figure 4 The welcoming screen of the website.

Stimuli Subjects were given nine different concepts to describe, split into difficulty levels and noun types (a total of 18 descriptions per game). We used the predicates most used in OMCS, indicating they are the easiest to describe (e.g. HasPrerequisite (*what do you need to do first*) was left out). Templates should be used to describe noun phrases (NP), verb phrases (VP) as VP = NP were not of use since no verbs were in the wordlist (of course, children could use verb phrases in their descriptions). The six predicates used make for 70% of total assertions in the current English OMCS (with UsedFor and CapableOf together covering app. 40% (Liu and Singh, 2004)), indicating these are indeed easier and more relevant.

Relation-type	Question posed	Syntactic form
MadeOf	What is it made of?	NP is made of NP
IsA	What kind of thing is it?	NP is a kind of NP
UsedFor	What do you use it for?	NP is used for VP
CapableOf	What can it do?	NP can VP
PartOf	What is it part of?	NP is part of NP
AtLocation	Where would you find it?	NP is found at NP

Table 3 Predicates used in the experiment

Stimuli were selected from the "Woorden in het basisonderwijs" (Words in primary school), developed by Schrooten and Vermeer in 1994. The list consist of over 26,000 words collected from different corpora presented to children in primary education, indicating which words are expected to be acquired at a given age and how large the lexicon of a child will be.

The words are devised based on grade (class) in which they are presented and grouped on their lemmas, meaning an occurrence of a word is collected as its basic form. With the Schrooten and Vermeer-corpus we have a reliable base for selecting words our subjects are familiar with. For the wordlist, only nouns were selected (which are easier to describe). Combined lemmas ("pictureframe") we are excluded. We applied a certain category filter, since random selections from the list gave a lot of animals and bodyparts - whilst it is certainly relevant to know you'll find fish near the sea and deer near the woods, for the purpose of our research and analyses we hand-picked just of few of those type of categories (as with house, barn, building, garage and synonyms in general).

Selected nouns were split into difficulty levels by measuring the word frequency (how often a word appears in a corpus) and spread (whether the word appears very often in only one text or often in multiple texts) in Schrooten and Vermeer's word list . The more a word is presented to a child, the sooner that word is acquired, thus the lower assigned category.

- | | |
|------------------|--|
| Easy | Words all subjects are expected to be familiar with. |
| Average | Words the median of subjects are expected to be familiar with. |
| Difficult | Words highest-grades are expected to be familiar with. |

For a technical description of assigning difficulty levels and thresholds I refer to Eckhardt (2008).

For the three difficulty levels, words were manually categorized as being abstract ("love") or concrete ("tree"). Concrete nouns refer to observable (physical) elements. Abstract nouns refer to non-observable elements. Main distinction between abstract and concrete in the context of our experiment is the level of difficulty: although "love" was defined an easy abstract concept it is much harder to describe than "kiss" (an easy concrete concept).

Pilot Study Before putting the experiment on 'live', a short pilot study, providing useful information concerning technical issues, motivational design and gameplay was held at a daycare center. In the pilot, six children, aged eight to twelve years played the online game (with supervision). Children were allowed to ask the supervisors any questions about the gameplay, but not about any of the words or answers.

Results from the pilot study showed children found abstract concepts hardest to describe (sometimes one word took them over five minutes; regardless of the option to skip a word). Following from our observations, it was decided to increase the number of concrete, easy stimuli (from 40 to 20), thus raising the probability an easy-to-describe concept will be presented to a subject, helping the willingness to keep playing (avoiding discouragement).

Procedures Children were pulled towards the website with the use of snail mail (see participants). On the welcoming screen a text-based description of the game and its context were displayed from the perspective of the robot and the character was introduced.

Players were asked to fill out some info on their characteristics relevant to our research - Name, Age, Email, Sexe, Class, Mother tongue. Age, Sexe, Class and Mother tongue were radio buttons (only one predefined choice from a list). Several children reported they were either younger or older and were told to pick the nearest age. It was then immediately corrected by hand in the underlying database and excluded from the analysis.

Instruction consisted of both a step-by-step visual explanation as well as text. Children received instruction both before the start of a phase and could also recall the information during playing. With the use of a roll-over image on every screen, children could get hints on what to fill out where at any given point during the game. Also, a link was provided to recall the original step-by-step instruction.

All instructions (welcoming screen, phases) were pre-tested in both a pilot study and by the expert eyes of two pre-school teachers. The foregoing invitation was also checked and marked by these teachers.

When playing, from a predefined list of concepts children were presented 9 concepts to describe (either easy, average or difficult for concrete or abstract concepts, one at a time). To describe them, they could choose from the predicate types used in OMCS (Table 2) in a drop down menu. Every predicate type could only be used once per concept. Descriptions were entered in a textbox. For every concept three descriptions had to be given.

To bring and keep motivation to contribute properly as high as possible, during every step it was communicated to children playing there where no strings attached - one could not give any wrong answers and it was stressed it was no intelligence test. All was aimed at avoiding research-effects as we wanted to know if the game design in itself was motivational enough, emphasizing children were participating in a research project would be disturbing ("what's research?") and impairing to those goals.

To elicit motivation within the participating children, we incorporated an element of competition. In the end, all players were rewarded with a second place in a top three as "second best teacher of Robot Rob" of the day. We explicitly chose the second place to encourage them to play it again ("want to be first") and to never let them down.

When a certain concept was found to be too hard, children had an "exit!"-option and were given another concept not to discourage them. From the pilotstudy it appeared abstract concepts were very hard. Therefore, the number of concrete (easier to describe) concepts was raised.

It itself, regarding the stages of development the children were in, it was expected the keenness on telling someone or something about what a child knows is quite motivational in itself. From comments it appeared this was indeed true. Nonetheless, the persona used was especially motivational in relation to the analogy with "teaching the computer" (computers truly are a

second nature to children), it gave a sense of a collaborative effort, a buddy with or for whom children were playing.



Figure 5 All children could print a certificate as a thank-you for contributing

4.1.2 Results⁶

Within the first three weeks after sending the letters to the children, the game had been played an approximate 150 times (the data from over 50 games were excluded from the analysis as false entries, such as demonstrations by the researchers or bad intent), and we declared the data collection finished. $N=123$ (a dozen children played it several times with one child playing it over eight times), 63 subjects were male, 60 were female.

A total of 4077 descriptions were collected. Manually judging the descriptions children provided (with 25% Simple Random Sampling) showed over 92% of all descriptions were reliable.

⁶ Please note the first experiment is mainly regarded a way to collect the stimuli for the second experiment.

4.2 Experiment 2

With the second experiment we wanted to evaluate if the proposed in-game validation proves to be a valuable approach to automatically collect commonsense with regard to truthfulness and whether or not (and to what extent) the number of cues (assertions) given are of influence to the number of correct answers.

4.2.1 Methods

Participants 119 ten to 12-year-old children in primary school (grades seven and eight), spread over four classes (two classes per grade). No specific efforts to control the consistency of the groups apart from in-class distributions was made (gender, age).

Design We used an independent measures design (between subjects). Correctness of the answers is used as the dependent variable, condition is used as the independent variable (measuring the effect of condition on correctness). The data were analyzed using logistic regression (with binary scoring)

Materials The questionnaires consisted of an introduction, brief example/explanation, questions and some supportive clues. Three versions were made (one for each condition), varying to the number of clues (assertions) given to make a guess. The questionnaires contained the assertions from the first experiment that belonged to 18 concepts from the database. One set of descriptions of one given concept is regarded one stimulus.

Children had to guess a concept based on a given number of descriptions (concept- predicate - description). All questionnaires were equal part from the manipulations.

The questionnaires are included in the appendices.

Stimuli The stimuli used consisted of the original descriptions (concept- predicate - description) minus the concept. Stimuli were presented per type (abstract vs concrete) and per difficulty level (easy, average, difficult).

The concepts on which the stimuli were randomly sampled (using Simple Random Sampling with equal probabilities in PASW Statistics 18) for every difficulty level and type (abstract-concrete) using a split data file. We did not recompute sampling probabilities consistently with the findings from the first experiment. We did not randomize between conditions, every child was presented the assertions for the same words since the proportion of our group of subjects was regarded too small and we were at risk of results in effect of the randomization instead of the conditions.

Difficulty	Concrete	Abstract	Total
Easy	3	3	6
Average	3	3	6
Difficult	3	3	6
Total	9	9	18

Table 4 Selection of descriptions from the original data per questionnaire

For means of validity the original data were not altered in any way, including typographical or grammatical errors. Selected concepts and descriptions were checked for recursive problems (descriptions consisting of or containing the original concept) by hand. No changes had to be made.

Example stimulus, (Condition clue3, e.g. Concrete level 1)

 ...is made of - wood,
 ...can - float,
 ...used for - transport

Conditions Conditions varied with the number of assertions (clues) per concept provided to the children, ranging from two to four descriptions . Every participant only received one version (condition) of the questionnaires. The condition was used as the independent variable.

When thinking of applying a validation component to the type of commonsense involved, it is very interesting to know to what extent the number of cues given in- or decrease the number of correct answers. If lowering the number of cues does not significantly impair the number of correct guesses (the reliability), removing the one cue can be used to validate more data with the same effort (raising efficiency while maintaining reliability). If the reliability significantly rises when more cues are presented, at the expense of efficiency reliability is also raised.

In the game, a standard stimulus in the second phase consisted of three predicate - descriptions. Two manipulations were carried out: from the nine concepts per level three were cut down to two descriptions to create a less-informed condition, three were left intact and another three were given a fourth predicate + description to create an augmented condition. For this fourth description an extra random sampled description from the database was used (belonging to the same concept).

\$clue2 - less specific

.....
...is made of - wood,
...can - float

\$clue3 - control

.....
...is made of - wood,
...can - float
... used for - transport

\$clue4 - augmented

.....
...is made of - wood,
...can - float
... used for - transport
... can be found near - harbor

Pilot Study Before conducting the real experiment the questionnaires and approach were tested in a small group of subjects. After it showed that guessing a concept based on one description was ridiculously hard it was decided not to use a one-description condition in the experiment. Apart from some children reporting the tasks easy and filling them out fast there were there were no shocking findings.

Procedures The experiment was conducted in classrooms of participating primary schools, the questionnaires were filled out as a type of exam by all the children in a class simultaneously. The researcher first explained what the idea was and children were handed out the questionnaires. Questions could be asked and instructions were read aloud before the children were given 10 minutes to fill out the questionnaires. Children were not allowed to communicate during the experiment. No efforts were made to make it any more fun than the task in itself was. Subjects received either one of the three conditions.

Scoring and coding All questionnaires were manually marked by the researcher, scoring the answers to seven categories in total, based on what occurred in the given answers. The correctness was used as the dependent variable. [table 5]

For the statistical analysis, scores were recoded into binaries, coding strictly not correct answers (Scores 3 to 7) as incorrect [table 6]

Correctness	Example	Score
	target > actual	
Correct		
One on one match with target concept (100% string similarity)	canoe > canoe	1
Wrong		
Plain wrong answers	cinnamon > cream	2
Correct, but with typo		
One on one matches with a misspelling	canoe > kanoe	3
Wrong, but semantically related		
A wrong answer that bears a semantic relation to the target concept	humor > clown	4
Wrong, but synonymous		
Basically a correct answer based on the description and very near the target concept.	fine > ticket	5
Wrong, but meronymous		
If the answer is an overspecification of the target concept (in Dutch usually a compound, e.g. soccer ball (<i>voetbal</i>)). Meronymy strictly also includes a more generic concept of the target concept, but it was not encountered.	flower > tulip	6
Wrong, but correct based on description	fine > jail	
If the the answer is not the target concept but correct given the assertions.	... <i>is a type of punishment</i> ... <i>can cost a lot of money</i>	7
Missing	?	9

Table 5 Scoring correctness of answers provided.

Statistics Since we want to know what the effect is of the number of assertions presented to a child on correctness of the answers they provide, the focus of our analysis is on the effect of *Condition* -independent- on *Correctness* -dependent-. More specifically, *how much* the value of correctness changes when condition is modified. Such questions are typically answered by running regression analyses (Allison, 1999; Hinkle et al, 1988).

For data where the dependent variable is binary or dichotomous, logistic regression is recommended (Wuensch, 2008). The other variables used in the questionnaires, word-level and -type, acted as control variables in the analysis.

A Binary Logistic Regression Analysis was performed, with *Condition* as the independent variable, measuring the predictive effect on *Correctness* as the dependent variable using PASW Statistics 18. The .05 criterion of statistical significance was employed for all tests (in earthlings language: we want to keep the chance of our findings being coincidence less than 5%).

Correctness	Score
Correct One on one match with target concept	1
Wrong	0
Correct, typo	0
Wrong, semantic match	0
Wrong, synonymous	0
Wrong, meronymous	0
Wrong, correct based on description.	0
Missing	9

Table 6 Recoding of the correctness-scores for the regression analysis. Please note correct-typo is regarded as wrong.

4.2.2 Results

119 questionnaires were filled out ($N=119$), $N=37$ for the less specific condition (Clue2, 31,1%), $N=42$ for the normal condition (Clue3, 35,3%) and $N=40$ for the augmented condition (Clue4, 33,6%). [table 9]

Of a total of 2142 stimuli, 658 were marked as correct, 1351 were marked as incorrect (30,7% vs. 63,1% respectively), 690 of which were plain wrong (32,2%), 6 were typos (0,3%), 140 were semantically related (6,5%), 138 were synonymous (6,4%), 55 were cases of meronymity (2,6%) and 322 were correct given the descriptions (15%). 133 were missing scores (6,2%). [table 8]

The distribution of correct vs incorrect answers over the conditions is showed in [table 7.]

Condition	<i>N</i>	Correctness	Frequency	%
(0) Clue2	37	Correct	191	28.7
		Wrong	446	67
		Missing	29	4.4
		Total	666	100
(1) Clue3	42	Correct	260	34.4
		Wrong	459	60.7
		Missing	37	4.9
		Total	756	100
(2) Clue4	40	Correct	207	28.7
		Wrong	446	61.9
		Missing	67	9.3
		Total	720	100
Total	119		2142	100

Table 7 Distribution of answers over conditions on binary scores.

Correctness	Frequency	%
Correct	658	30.7
Wrong	690	32.2
Correct, but with typo	6	0.3
Wrong, but semantically related	140	6.5
Wrong, but synonymous	138	6.4
Wrong, but meronymous	55	2.6
Wrong, but correct based on description	322	15
Missing	133	6.2
Total	2142	100

Table 8 Assigned scores.

Correctness	Frequency	%
Correct	658	30.7
Wrong	1351	63.1
Missing	133	6.2
Total	2142	100

Table 9 Assigned scores using binaries for regression.

A first indicator of the goodness-of-fit of the model used is the model summary. Here, indicates the logistic model suffices, where *Condition* explains somewhere between 25.5% (Cox & Snell \mathbf{R}^2) and 35.5% (Nagelkerke \mathbf{R}^2) of the variation in the *Correctness* [$P < 0.05$] [table 11]. That is confirmed by the Hosmer & Lemeshow-test [$P > 0.05$] (no significant differences between data in the study and as predicted by the regression model) [Tables 10 & 11]

Condition was of general significant predicative value to Correctness [$P < .01$]. [table 12]

Hypothesis

H1: The number of assertions used for guessing affects the correctness of the answers provided. - confirmed.

Condition(1) (three assertions) shows a statistically significant higher probability of correctness compared to Condition(0), [$B = .404$, $P > .01$] [table 12]

Condition (2) (four assertions) shows insignificant predictive value to correctness but indicates a slightly higher correctness [$B = .136, P > .05$] [table 12]

No significant interaction effects were found. The interaction of wordlevel by wordtype was insignificant, indicating no other variables but condition are of significant influence to correctness. [$B = -.353, P > .05$] [table 13]

An explanation of the tables can be found below.

Step	-2 log likelihood	Cox & Snell R^2	Nagelkerke R^2	χ^2	Sig.
1	1949.69	0.225	0.335	591.345	< .001

Table 10 Model Summary of the logistic regression analysis with pseudo R^2 .

Step	χ^2	Df	Sig.
1	5.945	7	0.538

Table 11 Goodness-of-fit (Hosmer and Lemeshow) of the logistic regression analysis

Step	Condition	B	df	Sig.	Exp(B)
1	(0) Clue2		2	0.009	
	(1) Clue3	0.404	1	0.003	1.497
	(2) Clue4	0.136	1	0.331	1.146

Table 12 Results of the binary logistic analysis

Step	Variables	B	df	Sig.	Exp(B)
1	level*type	-0.353	1	.220	0.702

Table 13 Interaction effects of level and type

Explanation of the tables

Table 10: The χ^2 -test compares the oddsratio of the predicted model (**-2 Log Likelihood**, 1949.69) to the oddsratio of a model with a constant instead of a variable (**Initial Log Likelihood Function**, in this case 2541.035). The difference between these is the χ^2 . The **R²** is the proportion of variability in our data the statistical model accounts for. Please note these **R²** 's are regarded pseudo measures for logistic regression (Sieben & Linssen, 2009). [Table 11]

Table 11: The goodness-of-fit indicates the fitness of the model by comparison of the frequencies in the data and the frequencies in the regression model. The insignificance ($P > 0.05$) indicates there are no significant differences between our data and the regression model and thus a good fitness.

Table 12: The column **Condition** refers to our independent variables, with Condition(0) as the reference. The **B**-value indicates the direction and size of the effect, **Exp(B)** indicates the oddsratio of the effect (the effect on slope of the regression graph). Values **B** and **Exp(B)** are empty since Condition(0) cannot use itself as a reference. The **df** (degrees of freedom) are 2 for Condition(0) because it is used as a reference to Condition(1) and (2). Condition (1) and (2) have 1 degree of freedom since Condition(0) acts as the reference. The **Sig.** indicates the statistical significance of the findings with a 95%-interval (.05-statistical criterion).

5. Conclusions

Do children prove a valuable resource for collecting commonsense?

In our first experiment, with over 92% of the descriptions the children provided being judged as reliable, we proved children aged ten to twelve years can draw upon a decently sized vocabulary to describe a word. On top of that, following from the number of children that played the game several times and the feedback we received, children like the task of describing words and providing the computer with commonsense. Thus, the first experiment indicates children are indeed a reliable source to collect commonsense from.

The second experiment, aimed at the idea of in-game validation, indicates guessing a concept based on the assertions provided by other children can contribute to automated validation of the descriptions, but the extent is relatively restricted. Of all the guessed words in the second experiment, 2142 answers, only 30% (658) were one-on-one matches with the target concepts, indicating 30% of the data can be validated automatically using 100% string similarity. Of all data marked as wrong, 63% (1351 answers), almost 16% (339) (scores 3 - 6; typo, synonymy, meronymy, semantically related) can also be validated but only by using secondary systems such as lexical databases. The most useful type of answer, score 7 (Incorrect, but correct given the descriptions, indicating it is another concept that fits the description), makes up for another 15% of answers but can never contribute to a commonsense database (since it would need the type of processing we need a commonsense database for to mark it as correct).

Subquestion *Can in-game validation be a means for automated acquiring of truthful commonsense using children?*

Automated validation is a valuable means for acquiring commonsense from children for 30% of the descriptions provided, checked on 100% string similarity. Combining the scores that are not strictly wrong and their frequencies, using secondary validation for the strictly not-correct answers, almost 61% of descriptions provided in a commonsense-type game like *Gezond Verstand* can be used for automated validation, not requiring any other human interference.

Manipulating the number of assertions in *Condition* showed that presenting the children with three assertions raised the correctness significantly over presenting them with two assertions. Presenting the children with four assertions showed a marginal and insignificant improvement of correctness over presenting them with two assertions. Especially the latter case is interesting given *clue4* contains double the assertions: somehow, regardless of more information on attributes of a concept, it turns out children are not able to improve their guesses.

Subquestion *Does the number of assertions used for guessing the target concept influence reliability?*

Yes, by statistically significant margin the reliability, by means of correctness, is influenced by the number of assertions provided in guessing a concept. However, this

only holds for presenting the children with three over two assertions. Presenting more assertions does not significantly improve the reliability.

Based on the reliability of the answers the children provided us with using four assertions, we can conclude there is no need to use four assertions since it does not contribute to a higher correctness and thus usefulness of the answers provided. The effort in validating can therefore be used for validating more concepts (compared to presenting children with four assertions for guessing). We cannot maximize the efficiency of the effort by presenting them with two assertions either, since this would impair the reliability of the answers provided.

Main Question *Can children (ages 10-12) be a reliable source in adding commonsense facts to a commonsense database, can their contributions be automatically validated and can we raise the efficiency of their efforts?*

Our experiments show children aged ten to twelve can contribute reliable assertions of commonsense to a commonsense-database and they are motivated to contribute and finish the task of describing. Almost 30% of their contributions can be validated automatically within the game without human interference. Raising or lowering the number of assertions used does not yield a higher reliability, thus the recommended number of assertions for both describing and guessing remains three. The effort the children make in guessing concepts cannot be maximized by presenting them with only two assertions.

5.1 Limitations

First of all, our findings are only valid within the age group of our subjects (ten to twelve years of age) and their stages of development.

The insignificant correctness of four assertions compared to two assertions is a curious case for which our experiments do not give a proper explanation. It is expected that adding an assertion takes the combined descriptions, and thus the possible answers, from generic to specific. E.g. a means of transport can be about everything moving, whilst adding wheels and handlebars restrict it to motorbikes or ordinary cycles. It could be the answers provided by children in the clue4-condition were strictly too specific or synonymous. However, that has not been part of the analysis.

Word-levels and -types were kept as a constant for all conditions. However, that does not mean, when they would have been manipulated, they will not be of influence at all. For instance, it might turn out the guesses on assertions from easy-level concrete-type concepts are much more reliable (higher correctness), indicating a much larger proportion of these concepts can be validated automatically. This might prove a valuable turn to take in further developments of the game since a larger proportion of assertions can be validated automatically than found in this study. It should thus be explored.

The main argument for conducting the experiment in real-life was we were able to completely avoid the problems that might have occurred in the online game: those of look-ups and ‘parental help’. The large difference in reliability between the first and second experiment, 92% vs. 30%, might indicate the children relied quite heavily on such secondary sources in the online game. We should, however, note the task of guessing is regarded a much harder one than describing, due to the size

of the set of possible answers (an almost infinite set of descriptions can be given whilst the number of words fitting the descriptions is much more restricted). A normalized number for both experiments in reliability cannot be computed due to the difference in the type of task; therefore a valid comparison cannot be made and the results per experiment can only be interpreted for its specific task.

The task of guessing words seems to be a tedious one, with a lot of possible outcomes leading to some children reporting tip-of-the-tongue problems. That is, semantically near concepts or synonyms did prime, but they were stuck with an annoying feeling, fully aware of their answers not being completely correct but unable to give the correct answer nonetheless.

While the concepts in the questionnaire were purposely not randomized to avoid research effects, ironically it might be that lead to a research effect. On scoring the questionnaires, it looked like there was remarkable agreement among synonyms (same synonym as answer to same stimulus) in the answers (e.g. fine was used a lot where it should have been ticket [bekeuring > boete]). Also, some assertions in the stimuli seemed to be generally distractive or even prone to prime certain (faulty) concepts (like [sand... can be found alongside the beach] where *grain* was the actual target concept), indicating some assertions can overrule others, regardless of the answer given being incompatible or even invalid with all assertions. Several given synonyms showed a tendency to occur more often within one class (raft > canoe), raising the question if certain concepts are bound to have a relation to recent in-class encounters. When manually checked, this indeed seems to be true since the class' teachers made the same mistake as the children in this context.

Regarding statistics, it could be argued the number of cases used (N was 119) was too small. This is supported by the minimum cases (subjects) per independent variable for the logistic regression Spicer (2004) recommends: at least 50 cases per independent variable (ranging from 37 to 42 in this study). Finally, in logistic regression analysis, missing cases (a total of 133 stimuli in our set of 2142) are excluded. Strictly, missing cases could be valuable information, e.g. if certain descriptions are too hard. However, on collecting the questionnaires in-class and when scoring, it seemed some subjects had so much missing cases it can only be due to a lack of motivation.

5.2 Future Research

What implications does that have for a redesign of our game and further research?

First of all, the game is a good method to collect commonsense from children. Children like playing the game and they provide us with reliable assertions. Secondly, there is no need to redesign the in-game validation since an alternative number of assertions does not raise reliability or efforts can be maximized. The game will be a useful addition to OMCS.

In OMCS, subjects describe concepts they come up with themselves; in our game these words are assigned, with children describing words they would have not thought of themselves as a positive side-effect. As a result, in redesigning the game and as an addition to the current OMCS Dutch,

some thought should be put in the choice of assigned words; preferably using the Schrooten&Vermeer corpus for children.

With the current findings 30% of the data are appropriate for in-game validation. Part of the data marked as strictly wrong in this study could however be made into usable data by the use of semantic or lexical databases such as Cornetto for Dutch (Vossen, 2006) or Wordnet for English. With the use of such databases, using the answers that do bear semantic relations or are synonymous and meronymous gets in the reach. Given our findings, we can then raise the proportion of the data that is usable with another 15% - and even higher since information on synonymity (etc) gives not only information on the truthfulness of the target concept but on concepts that also fit the same description. The extent to which such databases can be used should be investigated and is regarded beyond the scope of this research. With only .3% of cases containing typographical errors, linking a spellchecker to the validation-part does not seem worth the effort. However, we must not forget that such an augmentation could provide useful for the description part - though children apparently are not influenced by errors in the descriptions when guessing the target concept, commonsense-reasoning does rely on the correctness of the commonsense provided.

Also, a replication of the study, slightly altered to avoid the problems indicated in the limitations and with a (much) larger set of subjects, should be conducted. It should incorporate an analysis that not only indicates the effect of condition on (binary) correctness but also on the other scores and the effects of word-level (easy, medium, difficult) and word-type (concrete, abstract). These effects are regarded beyond the scope of this study and are only used as constants but might also yield interesting results. That is, however, a very complex exercise.

Finally, a re-analysis of the data using the words itself instead of scored categories could be conducted, possibly answering the question around synonyms and the in-class research effects with regard to agreement.

5.3 Final words

Commonsense is a fascinating problem that might just get us that little bit closer to developing a better relationship with computers and machines in general. It is a problem of great scale that most humans never even get to master throughout their whole life.

It is exactly for those aspects we could -and should- use humans to use computers' endless memory to collect all those facts we collect throughout our lives - with all our efforts combined it is unlikely we will miss out on the many facts we miss out on individually.

However, that is a long way to go still, finding ways that can bring us a little bit closer to solving the problem of commonsense all contribute to that goal. We proposed a way of collecting commonsense using those that are right in the middle of acquiring it, children, looking for ways that make it as fun as possible and require as little interference and monitoring of *other* humans as

possible It shows a valuable turn to take in search for all the commonsense available (for dutch, at least).

If the game proposed were to be used next to OMCS-Dutch, incorporating the in-game validation, and aimed at both adults and children, a varied and broad collection of useful commonsense knowledge could be gathered and maybe, over the course of years, we could get a lot closer to building a commonsense database that can be used for actual, human-like, machine-reasoning.

7. References

- [Allison, 1999] Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.
- [Atherton, 2008] Atherton, J. S. (2008). *Learning and teaching: Piaget's developmental theory*. <http://www.learningandteaching.info/learning/piaget.htm> Retrieved 29/11/2009.
- [Ball et al, 1997]. Ball, G., Ling, D., Kurlander, D., Miller, J., Pugh, D., Skelly, T. et al. (1997). *Lifelike computer characters: The persona project at Microsoft research*. *Software agents*, 191-222.
- [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.
- Bumgardner, J. (2007). *The Origins of Mindstorms*. *Wired Magazine*, march 2007. Retrieved december 16th 2009 from http://www.wired.com/geekdad/2007/03/the_origins_of_/
- [Chung, 2006] Chung, H. (2006). *GlobalMind - Bridging the Gap Between Different Cultures and Languages with Common-sense Computing* MS Thesis, August 2006.
- [Eckhardt, 2008] Eckhardt, N. (2008). *A "Kid's Open Mind Common Sense": the relation between common sense development and vocabulary growth and sense distinction in children*. Ma Thesis, Human Aspects of Information Technology, Tilburg University, 2008.
- [Gray, 2003] Gray, Jeff (2003). *Collection of Ambiguous or Inconsistent/Incomplete Statements*. <http://www.gray-area.org/Research/Ambig/> Retrieved 29/11/2009
- [Harel & Paper, 1990] Harel, I, & Papert, S. (1990). *Software design as a learning environment*. *Interactive Learning Environments*, 1(1), 1P32.
- [Hinkle et al, 1988] Hinkle, Dennis E., William Wiersma, & Stephen G. Jurs. (1988). *Applied Statistics for the Behavioral Sciences*. Boston: Houghton Mifflin Company.
- [Kuipers, 2004] Kuipers, B. (2004) *Making Sense of Common Sense Knowledge*. *Ubiquity*, Volume 4, Issue 45, Jan. 13 - 19, 2004
- [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.
- [Lester et al, 1997]. Lester, J. C., Voerman, J. L., Towns, S. G., & Callaway, C. B. (1997) *Cosmo: A life-like animated pedagogical agent with deictic believability*. Paper presented at the Working Notes of the IJCAI97 Workshop on Animated Interface Agents: Making Them Intelligent.
- [Liu and Singh, 2004] Liu, H. and Singh, P. (2004). *ConceptNet: A Practical Common-sense Reasoning Toolkit*. *BT Technology Journal*, 22(4):211-226.
- [McCarthy, 2007] McCarthy, J. (2007). *From Here to Human-level AI*. *Artificial Intelligence*, 171 (18):1174-1182.
- [McCarthy, 1959] McCarthy, J. (1959): *Programs with Common Sense*, Proceedings of the Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationery Office, London.

- [McCarthy & Hayes, 1969] McCarthy, J. and Hayes, P.J. (1969): *Some Philosophical Problems from the Standpoint of Artificial Intelligence*, D. Michie (ed.), Machine Intelligence 4, American Elsevier, New York, NY.
- [McKinstry, 1997] McKinstry, C. (1997). *Mind as Space. Toward the Automatic Discovery of a Universal Human Semantic-affective Hyperspace A Possible Subcognitive Foundation of a Computer Program Able to Pass the Turing Test*. In Epstein, R., Roberts, G., and Beber, G. (2008) *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. 1st. Springer Publishing Company, Incorporated.
- [Minsky, 1988] Minsky, M. (1988). *The Society Of Mind*. Simon & Schuster: New York.
- [Minsky, 2006] Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster: New York.
- [Moore, 2006] Moore, G. (2006) *Excerpts from A Conversation with Gordon Moore: Moore's Law* (PDF). Intel Corporation. 2005. pp. 1. Retrieved 29/11/2009.
- [Mueller, 2000] Mueller, Erik T. (2000). *A calendar with common sense. Proceedings of the 2000 International Conference on Intelligent User Interfaces* (pp. 198-201). New York: ACM.
- [Mueller, 2003] Mueller, E. T. (2003). *Thoughttreasure: A natural language/ commonsense platform*. <http://alumni.media.mit.edu/~mueller/papers/tt.html>. retrieved 29/11/2009
- [Van Mulken et al, 1998]. Van Mulken, S., André, E., & Muller, J. (1998). *The persona effect: how substantial is it?* People and Computers, 53-66.
- [Papert, 1980] Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books.
- [Schrooten and Vermeer, 1994] Schrooten, W. and Vermeer, A. (1994). *Woorden in het Basisonderwijs. 15000 woorden aangeboden aan leerlingen*.
- [Singh, 1999] Singh, P. (1999). *Acquiring common sense over the Internet*. MIT Media Lab.
- [Singh, 2002] Singh, P. (2002). *The Open Mind Common Sense Project*, MIT Media Lab.
- [Singh, 2003] Singh, P. (2003) . *Examining the Society of Mind*. Computing and Informatics, 22(5): 521-543
- [Singh & Barry, 2003] Singh, P. & Barry, B. (2003). *Collecting commonsense experiences. Proceedings of the Second International Conference on Knowledge Capture (K-CAP 2003)*. Sanibel Island, FL
- [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>. retrieved 29/11/2009.
- [Spicer, 2004] Spicer, J. (2004) *Making sense of multivariate data analysis*. Sage Publications, California, 123-151.
- [Stork, 1999] Stork, D. G. (1999). *The Open Mind initiative*. IEEE Expert Systems and Their Applications, 14:16–20.

- [Sieben & Linssen, 2009] Sieben, I. & Linssen, L. (2009). *Logistische regressie analyse: een handleiding*. <http://www.ru.nl/aspx/download.aspx?File=/contents/pages/451023/logistischeregressie.pdf&structuur=socialewetenschappen>, retrieved 29/11/2009
- [Valin & Stork, 1999] Valin, J.M. & Stork, D. (1999), *Open Mind speech recognition*, Proceedings of the Automatic Speech Recognition Workshop, ASRU99, Keystone CO, December 12-15, 1999.
- [von Ahn, 2007] von Ahn, L. (2007) *Human computation*, Proceedings of the 4th international conference on Knowledge capture, p.5-6, Whistler, BC, Canada
- [von Ahn et al, 2004] von Ahn, L., Blum, M. & Langford, J., *How Lazy Cryptographers do AI*. Communications of the ACM, February 2004. pp 56-60.
- [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 and Dabbish, 2008] von Ahn, L. and Dabbish, L. 2008. *Designing games with a purpose*. Commun. ACM 51, 8 (Aug. 2008), 58-67.
- [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, 2006] Vossen, P. (2006). *Cornetto: Een lexicaal-semantische database voor taaltechnologie*, Dixit Special Issue, Stevin
- [Wood et al, 2001] Wood, K. C., Smith, H., Grossniklaus, D. (2001). *Piaget's Stages of Cognitive Development*. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology*. <http://projects.coe.uga.edu/epltt/>, retrieved 29/11/2009.
- [Wuensch, 2008] Karl Wuensch (2008). *Bivariate Linear Regression*. Retrieved 28/11/2009 from <http://core.ecu.edu/psyc/wuenschk/docs30/regr6430.doc>