Constructing Semantic Knowledge Model based on Children Dictionary

Master Thesis

Masood Ghasemzadeh

Supervised by:

Dr.M.Rajman & Dr.J.-C.Chappelier

Artificial Intelligence Lab – EPFL University

2010 - Switzerland

Abstract

Natural Language is a rich source of knowledge. Processing natural language in a way that computer can understand and draw linguistic inferences from it is the main research track in the Natural Language Processing (NLP) field. Semantic Knowledge Model (SKM) plays an important role in modeling and representing different knowledge sources. In this domain, a project is defined to address the previous problems in SKM and propose a new solution to construct a new one. The whole research project is based on a hypothesis: to have an ideal knowledge model, we need to construct our knowledge model based on the available knowledge source (e.g. dictionary definitions), from a very basic level such as Children dictionary and then integrate and extend it to a higher level e.g. Adult dictionary. This master thesis tried to evaluate part of this hypothesis, the ability to expand the definitions in an automated way and its potential to have the same mechanism at higher level. *Expansion* is used to expand the definition i.e. replacing every meaningful word in it by its own definition. The expansion mechanism helps to gather the sparse knowledge in dictionary. In this direction, "Core words" are defined. They constitute minimal set of words from which all other words can be defined by, in the dictionary. Besides what have been mentioned about core words, they help to control how far a definition can be expanded (otherwise it could be expanded unlimitedly). The problems which we faced during expanding the dictionary definitions are analyzed and our suggested solutions summarized carefully. The license of definitions in the Oxford First Dictionary has been bought by AI lab at EPFL. Due to the time limitations, only the verbs in the children dictionary which correspond to 429 head-words were analyzed and expanded.

Chapter 1 provides some basic information about Natural Language Processing, our motivations to start such a project, review of the three similar projects and our goal in this research. Chapter 2 details knowledge sources which we considered. In this chapter, we also analyzed the structures of the dictionary definitions. Chapter 3 introduces *expansion* mechanism and the relevant concepts. In chapter 4, practical issues in the expansion of children dictionary definitions are reveled. In this perspective, all the problems encountered during expansion steps are listed, categorized and analyzed. Conclusion and the future work is stated in the last chapter.

Keyword: semantic model, knowledge representation, expanding dictionary definitions, core words.

Table of Contents

Abstract 1

1. Introduction 1

1.1. Natural Language Processing 1

1.2. Motivations and problems 2

1.3. State of the art 3

1.3.1. Similar projects 4

❖ WordNet 4

❖ Cyc Project 6

❖ FrameNet 6

1.3.2. Why still we need new knowledge models 7

1.4. Goals 11

2. Knowledge Model 13

2.1. Knowledge Source 13

2.2. Dictionary as a knowledge source 13

2.3. Dictionary Analysis 14

3. Expansion of definitions 19

3.1. What is expansion? 19

3.2. Non-expandable words 21

3.2.1. Terminals 22

3.2.2. Cycles 23

3.2.3. How to find terminals and cycles? 24

3.2.4. Fundamental verbs 27

3.2.5. Functional verbs 27

4. Practical aspects 29

4.1. How to expand 29

4.2. Problems in expansion 33

4.3. Methodology 49

5. Conclusion & Future work 51

References 54

Appendix A (List of Terminals) 56

Appendix B (List of Cycles) 58

Appendix C (Fundamental verbs with their definition) 59

Appendix D (Auxiliary verbs) 60

Appendix E (hand-coded functional verbs) 61

Appendix F (Head-words which added) 62

Appendix G (All head-words which their definitions have changed) 63

Appendix H (split head-words) 67

1. Introduction

1.1. Natural Language Processing

Natural Language Processing (NLP) refers to computer systems that analyze, attempt to understand, or produce text and speech as human does (Allen 2003). NLP has been studied since long time ago. There are mainly two motivations in it: First, the technical motivation for building intelligent computer systems such as machine translation systems, natural language interfaces to databases, man-machine interfaces systems, speech recognition, text & data mining and analyzing, information retrieval in general, question and answering systems etc. Secondly, there are cognitive and psycholinguistic motivations focusing on the process and mechanism that human beings use to process the language (Bharati 1995).

The scope of NLP usages is very broad. The following presents some NLP applications to show how they can be used in practice.

- **Machine Translation:** Machine Translation (MT) system is a typical application which its main goal is to translate text or speech from one language to other, using computer systems. Examples are: Eurotra for European languages (Bharati 1995) and more recent projects are CANDIDE by IBM and Google Translator (2007).

- **Natural language interfaces to databases:** Working with data and Data Base Management Systems (DBMS) is an important task which NLP could help to deal with. A typical example would be the ability for a user to interact with data in natural language form. LIFTER by Henrix (1978) and INTELLECT by Harris (1977) are some similar examples which were developed some decades ago (Bharati 1995).

- **Question and Answering (QA) systems:** The main goal of a Question answering (QA) system is to give Natural Language (NL) answers in response to Natural Language questions with the help of lexical resources. Depends on the complexity of the question and system, a question can be answered only by retrieving information from knowledge sources or doing shallow search on available lexical resources or it can use more advanced techniques such as logic-based reasoning, template-based approaches or statistics and machine learning methods. Early example of Question and Answering machine was LUNAR (1977) by Woods which its main focus was on answering to questions about the moon rocks (Bharati 1995). More modern and recent cases of logic-based QA systems are PowerAnswer (2006) and Senso (2007).

- **Speech Recognition:** Speech Recognition is to converting speech to text and also it can contain semantic understanding of the speech too. It is used in many fields especially in situations where the hands and eyes of the persons are busy or it cannot be used to perform tasks for any reason. (Englund 2004).

All of these applications have a kind of knowledge source. For example, in speech recognition, the voice is stored as signals (information) which will be used as knowledge source. Undoubtedly, today, text is the primary media which represent and transmit huge amount of information all over the world. Books, newspapers and magazines, dictionaries, encyclopedias, e-mail, instant messengers, weblogs, websites, forums and other similar text resources are available, which contain extreme amount of knowledge and information. Such large textual resources needed to be organized and managed to be able to be useful. To continue to make progress in textual-information management, we need to explore and develop systems which have more meaningful understanding of text. Semantic Knowledge Representation (SKR) is a way to model and show the available knowledge. Semantic Knowledge Representation is a field of study which concentrates on using formal symbols to a collection of propositioned, believed by some acknowledged persons. (Levesque. 2004).

1.2. Motivations and problems

Modeling and representing the available knowledge and drawing linguistic inferences are the general problem which is addressed in this master thesis. Although linguistic inferring is one of the main goals of this research, we are not interested in "reasoning" to draw inferences. What is more attractive for us is, we would like to explore what is inside the knowledge sources and extract (by any mechanism such as pattern matching) the desired knowledge. This statement is emphasizing on two key points:

1. Knowledge is *already* available inside the knowledge sources.
2. We need to explore and extract them in some way, not necessarily by reasoning method.

So what would be interesting – and also a good motivation - in this perspective is, how to model the knowledge sources? Before answering to this question, it would be good to know what an ideal model is.

An ideal model has two important features:

- It must be able to cover (and represent) all the knowledge inside the knowledge source in a good way.
- Facilitates searching and exploration mechanism appropriately.

The idea of building such a model leads some critical questions:

- Which knowledge source to choose?
- How about the size and complexity of the knowledge source?
    - Should we start from a very complex source and build it? If yes, how?
    - Should we start from less complex and integrate to more complex one? If yes, how?
- Regarding constructing such a model, every step should be done manual or automated? Which steps are not possible to be done automatically?

According to these questions, some researches have been done and tried to facilitate computer systems to overcome them. At syntactic level, good successes have been achieved but at semantic level, as we will see in section 0, works such as WordNet, Cyc Project and FrameNet, either could not be very successful, was very limited and not practical or was domain-dependent.

1.3. State of the art

Among similar projects which have been done, several terms such as *Semantic knowledge representation, Knowledge modeling, Lexical database, Lexical knowledge base, Commonsense knowledge base, Knowledge base, Lexical resource* have been used to refer to similar kind of systems with somehow the same purpose. What they have in common is, all of them try to model the available knowledge, for example by constructing a semantic network, and provide a mechanism to have access to them e.g. developing a software which illustrates the model. In this report, we use *knowledge model* to call the similar systems. As explained, most of these works intended to use a lexical resource to model and represent knowledge but what make them different are: The purpose of the system, the source of knowledge they used and the method which they modeled the knowledge resource:

- **Purpose:** Depends on the requirements of a project, the purpose would be different. Mainly knowledge modeling is used in Artificial Intelligence but later on it has been used in Philosophy, Psychology, Linguistic (Sowa 1992) ,Sociology, Biology, Business etc. The purpose of a knowledge model is not limited. It can be designed and modeled for a very specific goal too. For instance, one is developed to assist linguistic inferences, another one aimed to help word sense disambiguation etc. The purpose of a knowledge model affects how to choose source of knowledge and the modeling mechanisms.

- **Source of knowledge:** The source of knowledge could be different too. For example, a medical knowledge system which its goal is to help doctors to diagnose needs particular kind of knowledge sources which are bounded to any kind of medical information. Besides the content of knowledge sources, the structure of them is principal too. Assuming that we are restricted to textual information, a knowledge source could be structure, un-structured and semi-structured. Data which is formatted and stored in a pre-designed database and can be fetched and managed by a Data Base Management System (DBMS) is an example of structured source. Un-structured sources have not any kind of specific format or order. The text in a page of a book which often has no format and no structure can be considered as an un-structured source and semi-structured sources are those which are neither totally un-structured nor structured purely. Semi-structured sources may have tags or any sort of markers which structure the text weakly.

- **Method:** The most popular semantic knowledge modeling and representation is "semantic network". Semantic network has different types. For instance, six of the most common kinds of semantic networks are *Definitional networks, Assertion networks, Implicational networks, Executable network, Learning networks and Hybrid networks*. **"**What is common to all semantic networks is a declarative graphic representation that can be used either to represent knowledge or to support automated systems for reasoning about knowledge. Some versions are highly informal, but other versions are formally defined systems of logic." (Sowa 1992). Also it is important to mention that the purpose and the way which this model is going to be used later on, affect the way which model is made.

1.3.1. Similar projects

Among available similar projects, three of them sound interesting to us named *WordNet, Cyc project and FrameNet.* Different aspects of each of these projects are analyzed in next parts. Also, in addition to three mentioned factors, some other distinguishing parameters such as accessibility to the project, size of resources which they used and the way which models were made are discussed.

❖ WordNet

Description: Probably WordNet is the most common and widely used semantic knowledge models. Its development started around 25 years ago at Princeton University (Liu 2004a). According to its official website [1], "WordNet is a large lexical database of English ... Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser". Each word sense is considered as a separated node.

[1] http://wordnet.princeton.edu/

(Liu 2004a) believes one of the reasons for its success and widely usage is because ,compare to other semantic knowledge models, WordNet is easy to use. Also, since it is a simple semantic network with words at the nodes, it is easy to use it for query expansion, or determining semantic similarity and other purposes.

Purpose: The purpose of WordNet is quite general. Some explanations are mentioned in (Fellbaum 1998) why they started to develop such a model. For example, in 1985 it was somehow obvious that relational lexical semantics is one possible alternative for componential lexical semantics and at the time, there was no well-developed relational semantic network. Another reason was, many researchers in computational linguistics and artificial intelligence wanted to work on knowledge representation and reasoning but apparently there were no complete lexical databases which would include word meanings as well as word forms which could be accessible via computers (Fellbaum 1998).

Resources: The resources for WordNet are various and they gathered their definitions from different resources. The list of resources which they used during past 25 years:

- Kučera and Francis's Standard Corpus of Present-Day Edited English (familiarly known as the Brown Corpus).
- Laurence Urdang's little "Basic Book of Synonyms and Antonyms" (1978).
- Urdang's revision of Rodale's "The Synonym Finder" (1978).
- Robert Chapman's 4th edition of "Roget's International Thesaurus" (1977).
- list of words compiled by Fred Chang at the Naval Personnel Research and Development Center.
- COMLEX , Ralph Grishman and his colleagues at New York University.

Method/Structure: Each node which is equal to a sense is linked to another node with some semantic relations such as “IS-A” or synonymy relation (Liu 2004a). Both nouns and verbs are organized into hierarchies, defined by “hypernym” or “IS-A” relationships.

Number of entities: In the latest available version (version 3.0) Roughly 207,000 word-sense pairs exist.

Accessibility: Online and downloadable. Free to use for research purposes and commercial usage under WordNet license.

Usability: Easy to use. Some resources and documents are available.

Data model built by: Totally hand-coded.

❖ Cyc Project

Description: The Cyc project was founded in 1984 in the Microelectronics and Computer Technology Corporation (MCC). The Cyc knowledge base (KB) tries to formalize commonsense knowledge into a logic framework (Lenat 1995). To have “reasoning” using Cyc, first it is needed to map the text into its logical representation which to do this, we need to use a specific language called CycL. But, this procedure is complex since all of the inherent ambiguity in natural language must be resolved to produce the unambiguous logical formulation required by CycL.

Purpose: Apparently, its first and main goal was to cover and model *common sense knowledge* (Lenat 1995). Also it could be useful in machine translation and speech recognition systems (resource: cyc.com).

Resources: Couldn’t find any information.

Method/Structure: Sets of assertions and many "microtheories" which each contains assertions that share a common set of assumptions.

Number of entities: The knowledge base contains 47,000 concepts and 306,000 facts (resource: cyc.com).

Accessibility: Two versions, OpenCyc is open source version of Cyc but limited. The complete version is only available for researchers with research purpose.

Usability: Difficult to use for reasoning tasks (Liu 2004a).

Data model built by: Totally hand-coded.

❖ FrameNet

Description: Initiated at Berkeley University, FrameNet is an on-line lexical resource for English, based on frame semantics. Each semantic frame can be considered as a concept which has a script. Generally it is used to describe things like an object, event or state. The way which they construct FrameNet is, they select words with special meanings, describe the frames or conceptual structures which underlie these meanings, examine sentences in corpus which have these words and finally record the ways in which the components of the sentences containing these words express information about the frames they evoke. (resource: framenet.icsi.berkeley.edu)

Purpose: The particular aim of this project is revealed by its official website as “The aim is to document the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, through computer-assisted annotation of example sentences and automatic tabulation and display of the annotation results." (Ruppenhofer 2010).

Resources:

- **Definitions of lexical units in FrameNet came from either:**
  - Concise Oxford Dictionary, 10th Edition (courtesy of OxfordUniversity Press) or
  - Definition written by a FrameNet staff member.

- **Corpora/Corpus which they used:**
  - In FrameNet I: British National Corpus
  - In FrameNet II: British National Corpus and North American Newswire corpora

Method/Structure: Concepts are expressed as frames. Each frame has a description which describes the frame. Also, each frame has a number of core and non-core frame elements which they act as semantic roles. Besides the frame, every lexical unit is associated with some other frame elements through annotations.

Number of entities: 11,600 lexical units which 6,800 of them are fully annotated - 960 semantic frames - 150,000 annotated sentences.

Accessibility: Online search is available plus a Frame Grapher which visualize the relation. Also two commercial and non-commercial (free) versions are available for download.

Usability: User needs to know the concepts and fundamental of Frames to be able to use it. Many user manual are available.

Data model built by: The only available information is showing that the annotation procedure was started and handled manually but some efforts have been made to make it fully automated for future (Fliedner 2004).

1.3.2. Why still we need new knowledge models.

Evaluation of these semantic knowledge models is not an easy task. To our knowledge, there is not any measuring mechanism or any scientific paper which compare different semantic knowledge models properly. But, reviewing different publications in this track is showing that there are some important factors for a semantic knowledge model which cause a knowledge model be considered as a successful model or not. However, when we are talking about 'being successful', we are talking very generally. For example a typical measurement to see if a model is successful or not would be, the rate of its usage among scientists though as we have seen, for cases such as WordNet, limited number of choices of knowledge model some decades ago, made (or perhaps forced) researchers to focus on one specific one. The important factors which we found in our study to assess if previous works fulfilled all requirements for active scientists in NLP are categorized into four classes':

1. Purpose
2. Source of knowledge
3. Usage
4. Data model construction : "Hand-coded" or "Automated"

Each of these factors is shown in Table 1 for different semantic knowledge models.

KM/Factor	Purpose	Last Ver.	Source	Usage	Hand-coded/Auto
WordNet	General NLP [1985-now]	• Ver. 3.0 (2006) For Unix/Solaris etc, : 155,287 words -117,659 synsets - 206,941 word-sense pairs • Ver. 2.1 (2005) For Windows	• Kučera and Francis's Standard Corpus of Present-Day Edited English. • Laurence Urdang's little "Basic Book of Synonyms and Antonyms" (1978). • Urdang's revision of Rodale's "The Synonym Finder" (1978). • Robert Chapman's 4th edition of "Roget's International Thesaurus" (1977). • list of words compiled by Fred Chang at the Naval Personnel Research and Development Center. • COMLEX , Ralph Grishman and his colleagues at New York University.	Arguably the most widely used	Hand-coded
Cyc	Common-sense [1984-now]	OpenCyc: • Ver. 2.0 (2009) [No exact info about number of entities] • Ver. 0.9 (2005) : 47,000 concepts - 300,000 assertions	Couldn't find any information	Needs special skills to use	Hand-coded
FrameNet	"semantic and syntactic combinatory possibilities of each word in each of its senses" [?-now]	• The last version contains 11,600 lexical units - 960 semantic frames exemplified in more than 150,000 annotated sentences	• American Newswire corpora • British National Corpus • Concise Oxford Dictionary, 10th Edition • Their staffs	Quite popular	Both

Table 1. Comparison of different factors in semantic knowledge models

The purpose of WordNet is not focused but Cyc and FrameNet both have particular goals. However claiming that they cannot be useful for general aims is not very smart since they are not domain-dependent and one might use them in Question & Answering systems efficiently but what is obvious, Cyc and FrameNet are limited. Other example of single purpose model is CoceptNet which is designed very precisely for specific reason. ConceptNet (Liu 2004a) is one of the noticeable work which is established at MIT media and it is constructed fully automated from Open Mind Common Sense (OMCS) corpus and its main purpose is to cover common-sense knowledge.

The second factor is resources. The resources which they used are quite challenging. Having different kinds of knowledge sources makes two problems:

1. How reliable is the knowledge in these sources?
2. Do we have integration in our model if we use different sources?

For example, in FrameNet some of the descriptions for frames are written by their own staffs. Without any doubt, high knowledge in lexicography is needed to define a word or phrase and proofing the validation of knowledge sources must be considered very carefully. Or when many various resources have been used during past 20 years in WordNet, how one could be sure that the knowledge which is inside the semantic model is integrated. Using different resources might lead some difficulties. Different dictionaries often define the same concept in very different ways (Atkins 2008). Let us have a look at an example. The verb “touch” is defined in different dictionaries by different publishers as follow:

Chambers (Primary Dictionary):

1. *To put your hand or fingers on something*.
2. *To make contact with a thing or person*.

Bounty Books (The complete dictionary & thesaurus):

1. *To feel with your hands.*
2. *To be against something.*

Collins (Primary dictionary)

1. *If you touch something, you put your fingers or hand on it.*

USBORNE (illustrated dictionary)

1. *To make contact with something, using your hands or other areas of your body.*

2. *To make gentle contact with another object.*

Oxford (First dictionary)

1. *If you touch something, you put your hand or fingers on it.*

2. If things are touching, they are so close there is no space between them.

As it is clear, each of these definitions defined the verb “touch”, but some emphasize on “put” (physical action), one emphasizes on “contact” and the other on “feel”. Also it is true for the *object* which a person uses to touch something else. According to these definitions we can touch with “hand”, “finger” or both. Interestingly, one definition could be considered as a more general form of the other one. The first definition by USBORNE is revealing that touch can be done by “part of your body” (includes fingers, hands, elbow etc). OXFORD defines touching as a special state which there is no space between two things. One can interpret “putting fingers on something” as “when there is no space between fingers and something”. So as we can see, choosing the most proper resource is very important since we need to have a reliable knowledge source and if we use different knowledge sources, we should consider integration and coherence between them. This important factor is not considered well in almost all the three knowledge models.

Another aspect which is important in evaluating a knowledge model is its usage. By usage we mean how much they have been used in different scientific works. To have a very realistic and simple view, when a system has high quality and could fulfill researchers’ needs, it will be used a lot. But sometimes, due to the lack of any other available options, researchers had to focus on the few available systems. WordNet is a good example of such a case. Miller, the founder of WordNet explicitly mention that in 1980’s there was a high demand for a semantic knowledge network based on relations and the main motivation for them to start such a project was lack of such a system at the time.

According to these facts, we believe that systems such as WordNet, Cyc and FrameNet are not silver bullets and some available criticisms support our belief that they cannot satisfy all needs of researchers thoroughly. For instance, Schubert tried to analyze and measure the quality of WordNet (Schubert. 2001) and they came to the conclusion that “from a uniformly distributed sample, we estimated that about 35% of WordNet’s hyponymy links do not represent a subsumption relationship between two predicates”. Additionally, WordNet could have been designed better. There are some points which are missing in WordNet. For example, it has formal taxonomic knowledge that “dog” is a “canine”, which is a “carnivore”, which is a ‘placental mammal’; whereas it cannot bring a practical member-to-set association which “dog” is a “pet” (Liu 2004a). Liu also explained that since WordNet mostly emphasizes on lexical aspects and used a formal taxonomic approach to relates words to each other (as ‘dog’ is-a ‘canine’ is-a ‘carnivore’) thus WordNet is suitable for word similarity detection and lexical categorization. The same type of problem is true for Cyc project. The main concentration of Cyc is on *commonsense* – which is special kind of knowledge - and it expresses them in a formalized logical framework. It is quite successful in careful deductive reasoning and its productivities are admirable appropriate for situations which can be expressed accurately and unambiguously (Liu 2004a) but difficulty to use Cyc for reasoning tasks plus the limited free version of Cyc caused it to not be considered as the first option for most textual understanding tasks (Liu 2004a).

To have a fascinating output in Cyc, what is necessary is a precise and well defined input (what which should be done with high focus) but as Gelernter (Gelernter 1994) emphasizes, human reasoning is not a zero-one process, it is a spectrum of mental focus. When the focus of person who is reason is high, logical and rational thinking happens. Traditional AI considered reasoning only when the focus of a person is extremely high but Gelernter adds that not majority, but some of human reasoning happens at low or medium focus and this is in contrast with Cyc conditions.

FrameNet is designed based on the frame theory. In frame semantics theory, a *frame* represents a scenario that contains an *interaction* and its *participants*, in a way that participants play some kind of roles. But according to (Shi 2005), "FrameNet does not explicitly define selectional restrictions for semantic roles".

The last but not the least is the way which these models are built. What is very obvious among available knowledge models is, between "hand-coded" and "automated" models there is a trade-off. Many efforts have been made to construct automated knowledge models by using pattern matching, statistical methods, information retrieval approaches, machine learning algorithms etc. e.g. (Collins 1975; F Suchanek 2007; Fellbaum 1998; Gelernter 1994; Lieberman H 2004; Liu 2004) but (Liu 2004a) point out that although the quality of automated models improved , the quality is still significantly below that of a hand-coded knowledge models. Also, he believes that having a recall above 90 percent for a closed domain typically causes a drastic loss of precision in return. By this, he expressed that automated models are only of little use for applications such as automated reasoning that need near-perfect ontologies. Moreover, he stated these models often do not have an explicit (logic-based) knowledge representation model. On the other hand, hand-coded models are extremely time and labor consuming. Also, high cost for assembly and quality assurance, limited coverage and updating are problems which these systems must deal with. None of the hand-coded models can be considered as the last and complete knowledge model.

All the mentioned problems are good motivations to think about new semantic knowledge model which covers a broad domain of knowledge, is automated but accurate and complete enough, by using reliable and integrated resources. This goal with its specified conditions formed the skeleton of the current master thesis project.

1.4. Goals

Related to the mentioned problems in the previous section, we have a solution for them. This solution is revealed in the form of a hypothesis.

- **Hypothesis:** To have a an ideal knowledge model, we need to construct our knowledge model based on the available knowledge sources, from a very basic level such as Children dictionary and then integrate and extend it to a higher level e.g. Adult dictionary.

There are four key points in this hypothesis:

1. Dictionary definitions are our source of knowledge and we will construct our model based on the dictionary. [detailed information why dictionary is used is discussed in section 2.2]
2. It is a data-driven model. It means, the knowledge resource will not be adapted to the pre-defined model, but in the other way, the model must be constructed according to the available knowledge resources.
3. Expanding the definition is one of the main procedures to build our knowledge model. [detailed in chapter 3]
4. We believe, while the level of the complexity of language increases (from children dictionary to adult), less manual work and time is needed to extend the model and it will converge to a roughly fully automated procedure meanwhile.

The main goal of this master thesis is to study the feasibility of the hypothesis. To reach this goal, we chose dictionary definitions as the raw data and apply the hypothesis on it. This process contains analyzing the dictionary definitions at children levels and represents them in a way that it can be used for:

- ❖ Semantic Knowledge Representation. (Short term goal)
- ❖ Linguistic Inferences. (Long term goal)

The first part contains the analysis of the definitions, finding a mechanism to expand the definition and finding the available semantic templates in the dictionary definitions. Due to the time limitation, only analyzing the definitions and expansion mechanism have been considered thoroughly in this master project. Available semantic templates in the dictionary and linguistic inferences postponed for the future works. Therefore, the main purpose of this master project can be defined as:

To analyze the dictionary definitions (as knowledge source) at children's level and develop a mechanism to expand the definition in a way that it supports the above hypothesis.

2. Knowledge Model

2.1. Knowledge Source

The textual format is a very flexible way to describe and store different types of information and large amount of information is stored and formatted as text. Therefore, such textual data is a valuable source of knowledge to explore (Nagano 2001). Among the available textual sources, we are interested in dictionaries. Next section describes why dictionaries were chosen as our primary knowledge sources.

2.2. Dictionary as a knowledge source

Knowledge can be thought of as concepts and the relations between these concepts (Chierchia 1988). A dictionary comprises of headwords and their definitions. Commonly each definition contains other headwords too. Depending on the type of the dictionary, it holds the definitions of words which people often use them in a particular language. They are good resources of knowledge. Besides, it is a reliable resource to refer to since the content of them are often generated by professional people which try to define each word in the best and accurate way so the facts which will be extracted from them would be reliable as well.

Despite the fact that dictionaries definitions are not formally structured, they often are well-defined and well-formatted. According to our observations, most of the definitions in a standard dictionary have some specific patterns. For example as it is shown in the Table 2, around 92% of the definitions in Oxford First Dictionary start with “If” or “When” and more than 73% of the definitions which start with “If” are followed by “You”. Although we are not sure whether all dictionaries and lexicographers use some limited and predefined templates in their definitions or not, many researches (Markowitz 1986; Smith 1981; Judith Markowitz 1986) made such an assumption. Thus, we assume it for our work too. A future work would be to prove such an idea and, if it is not the case, to study if it is possible to re-format them into the proper pattern.

However, there are some restrictions for dictionaries too. For example, adult dictionaries cover many broad and vast knowledge domains and it is often essential to have some previous knowledge or background to be able to understand adult dictionary definitions. This phenomenon sounds normal and logical because basically people get some basic knowledge from their childhood and while they grow up, their knowledge expands and becomes more complex accordingly, based on the previous knowledge. Let us have a look at an example. Below the definition of the verb “believe” by two dictionaries, same publisher but at different age group are shown.

Oxford First Dictionary (Age +5)

Believe: *If you believe someone or something, you think what is said is true.*

The new shorter Oxford (adult)

Believe: 1) *Have confidence or faith in or on (a person, God etc)*

2) *Put one's trust or have confidence in (or on) the truth of a (proposition, doctrine, etc)*

To understand what is "believe" in the second dictionary, one should know what is "faith", "confidence" or how is it possible to "put trust" on something or somebody, because for a child, "put" at first look means "the movement of something and leave it there" [physical]. The key point is that, when the level of the dictionary increases, the complexity of them goes up too. Probably it is at the higher levels (regarding the age group) which a child learns that (s)he can put something which is intangible (such as time, trust, energy etc) on the something or somewhere. Undoubtedly, processing such huge information (especially when it is about human natural language with its complicated features e.g. idioms, ellipsis, anaphora, paraphrase, ambiguity, vagueness, aspect, lexical semantics, the impact of pragmatics etc.) needs an exact and wise approach. By choosing the children dictionary, we can start from a less complex with few words. Moreover, it helps hand-coded processing a lot. Having fully automated model is an iterative process in our work and we believe initial steps must be taken manually in a small domain like children dictionary, and meanwhile increase the complexity of the model as long as the dictionary level increases by an automated procedure.

To follow such a strategy, we considered different dictionaries at different levels by different publishers. Each level contains different number of words and corresponds to different age group. For instance at first level, it will focus on the Children Dictionary which has around 2000 words. The whole proposed framework will have 3 levels: 2K, 6K and 20K words, Children, Teenage and Adult level respectively. Different dictionaries by different publisher will be used in each level since it helps to cover wider and more complete range of the words as well as definitions. Next section explicates the children dictionary definitions analysis and its results.

2.3. Dictionary Analysis

To create a knowledge model of the dictionaries definitions and have linguistic inferences, analyzing and having a comprehensive understanding about the dictionary definitions is crucial.

Important issue in the design of such a system is representation and integration of the acquired knowledge within the knowledge resources. The knowledge, which is available in dictionary, should be transformed into such a model which enables computer to make suitable integration of it, so the model must be close to human natural language. Additionally, the knowledge should be modeled in a way that facilitates making the plausible inferences. Moreover, the system should understand the questions and be able to find out the required knowledge. All of these conditions mean the knowledge model should be well structured (probably not exactly the same as human natural language but similar) so that computer can understand it (Kazimierczak 1990).

In this direction, we analyze the raw data to see how they are structured. Structure in definition could be syntactic or lexical templates which lexicographers use in a consistent way when they define a word in dictionary (Dolan 1993). For example majority of definitions have two parts which they are separated by a “comma” sign. Markowitz et al. and Smith (Kazimierczak 1990; Smith 1981) worked on the same idea which is a general mechanism by attempting to discover “defining formulae” or “significant recurring patterns” in the dictionary definitions. This step comprises of very low level text analyzing of dictionary definitions to find out some patterns or defining formula in definitions. Some of the analysis have been done manually and the other one automatically.

The input data got from ‘*Oxford First Dictionary*’ compiled by Evelyn Goldfish, Andrew Delahunty published by *OXFORD University Press* 2007. All the raw data is represented in XML format and XSLT language is used to analyze it.

To have a good order of words and be able to process them manually, all the available “verbs” at Oxford first dictionary have been chosen and explored. The very first analysis of dictionary definitions is shown in table 2.

Definition start with:			Number of nodes
<If> : freq="48%"	Followed by:	Percentage:	Total nodes : 206
	<you>	73.79%	152
	<something>	13.11%	27
	<someone>	8.74%	18
	<a> +>	1.94%	4
		followed by:	

		<person or thing>	(3 out of 4)
		<person or animal>	(1 out of 4)
	Other :	2.43%	
	<part>	<of your body>	1
	<one>	<thing>	1
	<people>		1
	<somebody>		1
	<things>		1
<When> : freq=44%	Followed by:	Percentage:	Total nodes: 190
	<you>	75.60%	143
	<something>	7.90%	15
	<people>	5.20%	10
	<a> +>	3.70%	7
		followed by:	
		<bird>	(2 out of 7)
		<person, animal, or plant>	(1 out of 7)
		<baby bird>	(1 out of 7)
		<person or animal>	(1 out of 7)
		<duck>	(1 out of 7)
		<thing>	(1 out of 7)
	<someone>		3
	<somebody>		2
	Other :	5.80%	
	<water>		2
	<liquid>		2
	<dogs>		1
	<things>		1
		followed by:	
	<two>	<people>	2
	<an>	<animal>	1
	<the>	<wind>	1
<To> : freq=7%	Followed by:		Total Nodes : 27
	25 different words		
Less than 1% (each) :			Total Nodes : 6
<An>	<animal>		1
<People>			1
<You>			1
<Words>			1
<Someone>			1
<Take>			1

Table 2.First, second and third word frequencies in Dictionary Definition [OxfordFirst 4-7-v1-100304 – Children level]

Table 2 shows the frequency of definitions according to their first and second words in their content. Totally, there are 335 verbs in the dictionary. It is possible that one verb has more than one meaning (sense). The total number of head-words is 429[2]. All the definition have been

[2] Head-word is any kind of word (in our case verb) which is defined in dictionary. A head-word is not necessary a unique verb. It can be different senses of the same verb which are defined separately in dictionary.

processed and categorized into four sub-groups. The groups are distinguished according to their first word of the definitions. For example, group 1, contains those definitions which start with “If”. Group 2 have definitions with “When” starting word. Third group contains definitions with “To” starting word and group 4 contains the rest.

Majority of the definitions start with either “If” or “When” which totally corresponds to around 92% of the whole 429 definitions. (Atkins 2008) explains the pattern of definitions like “If..” is a conversational format which helps the definitions to sound more natural in a way that it can be though as the answer to the question from user “What does this word mean?”. This technique is pioneered in COBUILD[3] (Atkins 2008). Our experience shows, it is possible to replace “If” with “When” (or the other way) in all definitions which start with “When” without any considerable changes in the meaning. It means, all the verbs which are defined by “When …..” can be defined by “If ….” as well (only the first word changed, the rest would be the same).

Also, what is more interesting for us is that, further analyze of group 1 and 2 expresses that around 98% of the definitions in group 1 and 2 have the same pattern in their definition and the remaining 2% can be changed in a way that they follow the same pattern. They are split into two parts by comma as below:

Definition: [*If|When* …] , […].

Such a high percentage revealing that the way which lexicographers defined the verbs is not random at all but most of the time (98%) verbs are defined with the same pattern. This is a well-known method in lexicography to define the definitions and it called *Full-sentence definitions (FSD). “*The FSD approach allows the lexicographers to embed these colligational and collocational preferences in the definition itself, giving learners a fuller picture of how the word is normally used” (Atkins 2008). More detailed information about FSD can be find in (Rundell 2006; Hanks 1987).

In FSD each definition is divided into two parts. In formulations of this definition type, the ‘left-hand side’ exemplifies usage, while the ‘right-hand side’ supplies the definition (Atkins 2008).

- Schema: The first part of each definition (left brackets). It often starts with “If” or “When” followed by, how to use the *verb* in a very simple form. For example, consider the definition of “Play”, *Play: When you play, you do something for fun.* This definition is split into two parts, before and after the ‘comma’. The first part is *[When you play]* and as we expect, it starts with “When” followed by a simple form of using the verb i.e. *you play.* We have named the first part as “Schema” part.

[3] A dictionary which is based on a "corpus" by "Collins Birmingham University International Language Database"

- Body: The second part tries to explain and define the *verb*. For example if we consider the definition of “Play” again we have: *Play: When you play, you do something for fun*. The second part (after comma) is explaining the definition of “Play” i.e. *you do something for fun.* We called the second part as “Body” part.

Additionally, the predictable patterns in the second word of the definitions are visible too. Around 94% of first word in verb definitions are followed by either <You>, <Someone> or <Something> which can be considered as a good evidence for a very simple patterns: Most of the definitions start with <If> or <When> followed by <Subject: You,Someone,Something>.

The *schema* part can be very short and simple or it can be complex and long. For sake of simplicity at Children dictionary, the *schemas* are often very simple and short. For example, the pattern ‘*If you <verb> , ...*’ have been seen in many definitions. Majority of definitions have subject or object as the second word in their content. According to the Table 2, distinctly, ‘You’ is the most frequent word which has been repeated in many definitions as the second word. In 73.79% of time, ‘You’ is the following word after *If* and 75.60% of times *When* is followed by ‘You’. ‘Something’ is the second most repeated word after *If* and *When*, 13.11% and 7.9% respectively. ‘Someone’ and ‘People’ are the third most repeated in Children Dictionary as the second word in definitions. This analysis not only showing that there are some basic patterns in the content of definitions but also it states that it is possible to change the *Subject* of the *Schema* part to make it more identical. For instance, in most of the case replacing *You* with *Someone* will not change neither the structure nor the meaning of a definition significantly. By doing such replacement, we make the definition format more identical and integrated. The *body* part is the most important part of each definition which contains valuable information and knowledge. Another usage of such analysis is, they would be very useful for future work for example how to find the reference of a pronoun in body etc.

3. Expansion of definitions

3.1. What is expansion?

Techniques to develop a dictionary has been studied for a long time and nowadays there exist many guidelines and recommendations how to write a dictionary efficiently. Defining a word in dictionary depends on many factors such as the user age and level and the purpose of dictionary etc but as a general view, (Atkins 2008) expressed some minimum needs for a good definition in a dictionary which must be fulfilled as much as possible. In his view one requirement to have intelligibility is: "The user shouldn't have to consult *another* definition in order to understand the one (s)he is looking up (this won't always be feasible, but it's a desirable objective)". But is it always true for all dictionaries? Theoretically it should be but what if this requirement is not fulfilled and a user needs to look up another definition to understand the first definition? Then one would say why shouldn't we do that for the user? *Expansion* mechanism tries to work out this problem.

In dictionary, a word is defined by some other words. We can replace these words by their own definitions. This is what we call *expansion* of a definition. An example can help to understand it better (The number in brackets shows the sense number):

Step 1) Bow[1]: <u>Bend</u> your head or body forward.

There is one verb, "bend" (underlined) in the definition of "bow" so what we will do in next step is to replace the verb "bend" with its definition. Below is the definition of "bend"

Step 2) Bend[1]: If you bend something, you <u>change</u> its shape so it is no longer straight.

By replacing its definition we will have:

Step 3) Bow[1]: you <u>change</u> your head or body's shape forward so it is no longer straight.

This is one simple example of expansion mechanism. We will see very soon why we didn't continue to expand the verb "change" in the last step.

At first look, the advantage of this mechanism might be its benefit for user to not look for other words in dictionary but it is not the only positive point for *expansion*. Besides, by expansion we are gathering different information which is available in dictionary, and put them in one sentence. Or, we can interpret it as, by expansion we describe a word in more detail. Previous expansion sample is a good case to show how expansion method provides more detail information about a word. Let us look at Bow[1] at step 1 and Bow[1] at step 3 together and see how expansion made them different.

Step 1) Bow[1]: <u>Bend</u> your head or body forward.
Step 3) Bow[1]: you <u>change</u> your head or body's shape forward so it is no longer straight.

The differences between these two definitions would be more valuable when we assume that a machine going to read and processes them. Imagine that a machine is supposed to try to extract information as much as it could do. By processing the first definition, what a machine can understand at first look is, by bow, you "bend" your head or body forward. Let's consider "understanding" as its simplest form. "Understanding" means making connection between facts. By this definition, a machine could find these connections for first definition (assuming that the machine will not process each connection any further):

"Bow" is connected to "bend"

"Bow" is connected to "head"

"Bow" is connected to "body"

"Bow" is connected to "forward"

And by processing the second definition, these relations are fetched:

"Bow" is connected to "change *head or body's* shape"

"Bow" is connected to "head"

"Bow" is connected to "body"

"Bow" is connected to "forward"

"Bow" is connected to "(not) straight"

What expansion is doing here is not increasing the knowledge or information in dictionary but in fact it is merging or putting together information which is spare in dictionary and it matches with the hypothesis very well. This mechanism could provide more information to machine. But the key questions in expansion mechanism are:

1. Can we expand all the definitions without losing or adding some non-sense information?
2. Does expansion alter the original definition a lot?
3. How much should we expand a definition? Is it finite?

The answer to questions 1 & 2 is addressed in section 4. Next section tries to consider the question 3.

3.2. Non-expandable words

To know how far can/should we expand a definition, we need to introduce some new concepts.

Non-expandable word: A "non-expandable" word is a word which when we see it in a definition, we decide not to expand it.

Core words: "core words" is *a minimal* set of "non-expandable word" from which all the other words in the dictionary can be defined by.

The definition for "core words" expresses two important properties which a set of "core words" needs to have:

a. It must be complete: Any other words in dictionary can be defined by this set.
b. It must be minimum: Generates and guarantee the uniqueness of expansion mechanism.

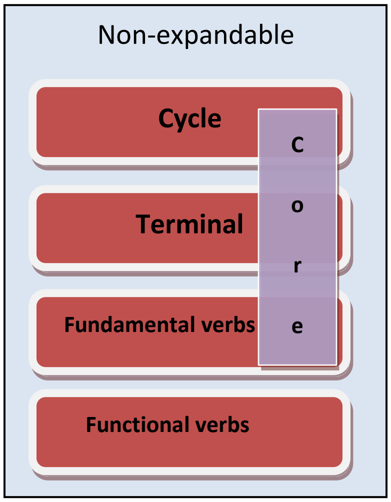

Figure 1. Non-expandable and core verbs

Non-expandable verbs have essential role in our work. According to the definition of non-expandable verbs, we categorize them in four different groups:

1. Terminals
2. Cycles (Loops)
3. Fundamental verbs
4. Functional verbs

A verb which belongs to any of these four groups is considered as a non-expandable verb. Definitions and the ways to find out each of these four groups are expressed in next sections.

3.2.1. Terminals

When we see one or more verbs in the definition of a head-word, we connect the head-word directed to the each verb. Finding the corresponding sense (Word Sense Disambiguation) has been done manually. For instance, the verb "fish[1]" is defined in dictionary as follow:

Fish[1]: If you fish, you try to catch fish.

In this example, the head-word is "fish[1]" and it is defined by two other verbs, "try[1]" and "catch[1]". We would connect them in our graph like Fig 2.

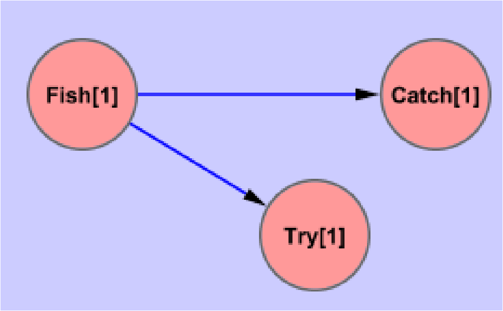

Figure 2. Connection between head-word and verbs in its definition

As we saw, if we consider the relation between a head-word and the verbs which are in its definition as a directed graph, then there are some nodes which have no out-going link. These nodes called "Terminal" or "Leaf". In other word, these nodes have no children in graph. But why we are interested in terminal nodes?

We are interested in terminals because a terminal means an ending point in graph. When a head-word is a terminal in dictionary it means we cannot define it anymore for any reason and it is exactly what we are looking for when we are dealing with the question "how far should/can we expand?"

When we are talking about dictionary, there are some few reasons to have terminals there. A head-word is terminal because either it is not included in the dictionary or it is defined by verbs which we do not consider them as a node in graph. The reason for second case is quite specific to our work and it is discussion in section 3.2 but there are some explanations for cases which a head-word is not included in a dictionary.

As far as we could find, a head-word is not included in a dictionary due to any of these three reasons:

- The lexicographer believes that this head-word is not necessary to be included in dictionary for linguistic reasons for example, the head-word is very simple and it doesn't need any description. We should be careful that there might be another reason to not include a head-word in dictionary by lexicographer i.e. the head-word is very complex for targeted user such as “metamorphosis” which is a complex word for a children. Lexicographers mainly try to not use complex words to define a word specially if there are some linguistic restrictions for users such as children dictionary.

- A head-word is not included in dictionary because of publication limitations. There are many restrictions for lexicographers while they are writing and developing a dictionary. Number of pages, size and font of content, density of words in each page etc are just some few examples of limitations which a lexicographer has. To get more information about how to develop and write a dictionary have a look at *The Oxford Guide to Practical Lexicography (Atkins 2008).*

- It is an error. The author forgot to include the head-word.

Besides finding the terminals, we need to have some strategy to deal with second and third cases. These issues are discussed in section 3.1. sub-part: missing-headword.

3.2.2. Cycles

Cycle or loop happens when word A is defined by word B, and word B is defined by word A. (Atkins 2008) recommends to prevent such a phenomena to happen as much as possible but sometimes it is not possible to prevent it. He used some examples from (*Newbury House Dictionary of American English*, 4th edition 2004) to show how a cycle would form. Let us have a look at definitions of “allow” “let” and “permit” below:

Allow[1]: to <u>let</u>; <u>permit</u>
Let[1]: to <u>allow</u>; <u>permit</u>
Permit[1]: to <u>allow</u>; <u>let</u>

What he emphasizes on is, some principles are too obvious and although we can define them but as we have seen in previous example some of them are practically irreducible. ”allow” perhaps is one of them. What he concludes is that a determination to avoid circularity at all costs may lead to definitions that are needlessly difficult. According to our experience, we can draw the same conclusion that when there is a cycle, probably words which are inside it are irreducible. When a word is irreducible, it is a good sign that this particular word is not expandable. Hence, when we find a cycle, we will not expand the words which are inside it but choose one of them based on a strategy which is discussed in 4.2.

3.2.3. How to find terminals and cycles?

To find terminals, we had to generate a connected graph and look for them there. The rest of this part gives an introduction about the connected graph and tools which we used to create our graph. Then the results of our exploration for terminals and cycles are mentioned based on our dictionary data.

Limited number of verbs which are 355 entities and 429 definitions lead us to decide to extract the verbs and disambiguate them manually. On average it took around one week for one person to extract all the verbs in the definitions and disambiguate them. After finding all the verbs, we form a graph. Two advantages of having a connected graph are:

1. We can visualize the connections and working with them would be handier.
2. By analyzing a connected network, we can extract very useful information from it such as finding strongly connected sub-graphs (cycles) and terminals.

The result of such a connected network which consists of 470 nodes and 587 directed edges (un-weighted) is shown in Fig 3.

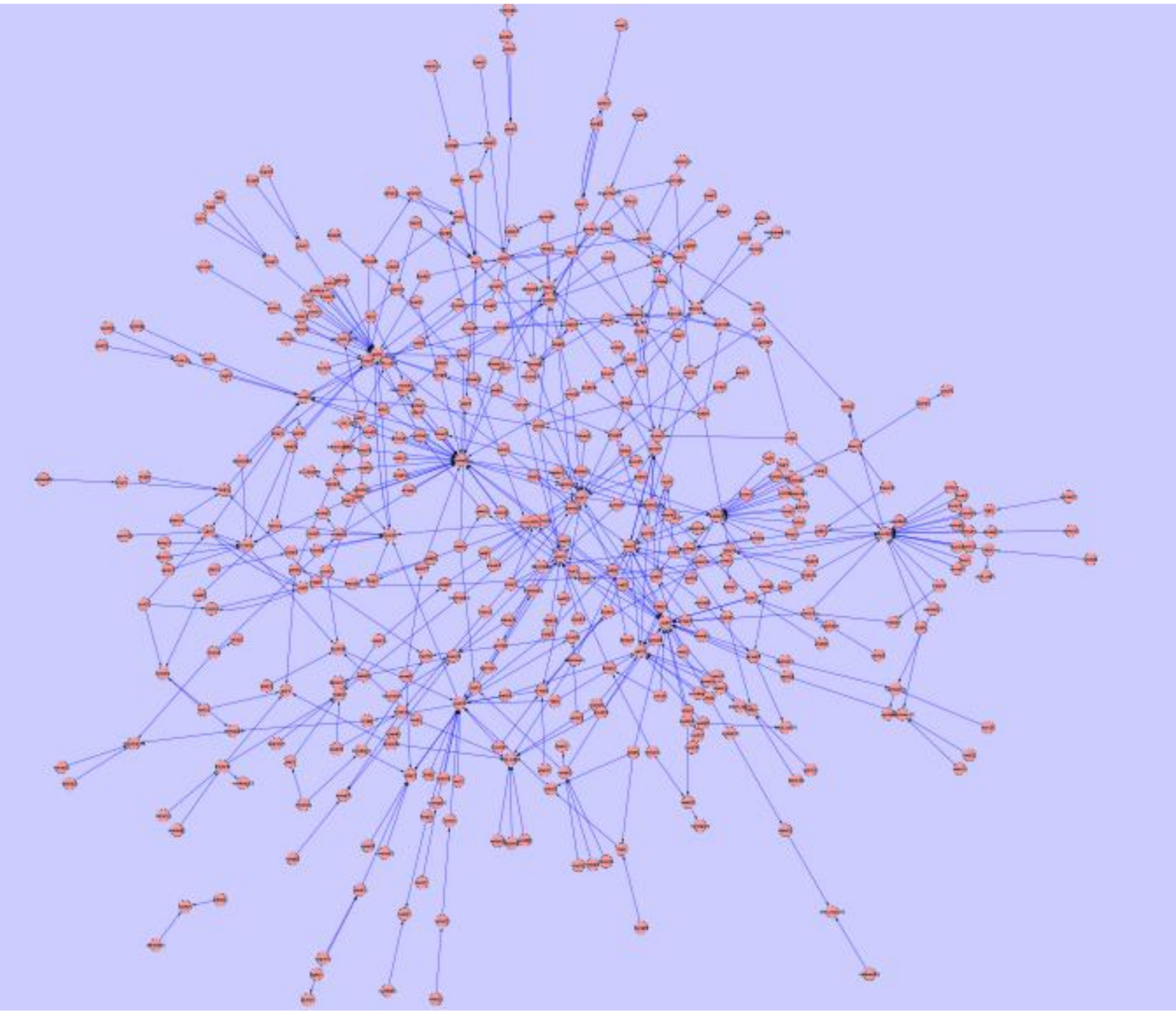

Figure 3. Connected graph of verbs included in definitions. Children level

Let have a look at some definitions examples and see how these nodes are connected to each other (Fig 4). Five verbs and their definitions are as follow:

Destroy[1]: If you destroy something, you damage it so much that it can no longer be used.
Damage[2]: If a person or thing damages something, they spoil it in some way.
Spoil[1]: If something is spoilt, it is not as good as it was before.
Burn[2]: If someone burns something, they damage it with fire or heat.
Scratch[1]: If you scratch something, you damage it by moving something sharp over it.

According to the fact that auxiliary verbs are excluded from our graph (see section 3), the verb 'Damage [1]' is used in the definition of 'Destroy[1]' so in our graph 'Destroy[1]' node is connected to 'Damage[1]' node. Similarly, the verb 'Spoilt[1]' is used in the definition of

‘Damage [1]’ so the node ‘Damage [1]’ is connected to ‘Spoil [1]’ and so on. The visualized graph of these connections would look like Fig 4.

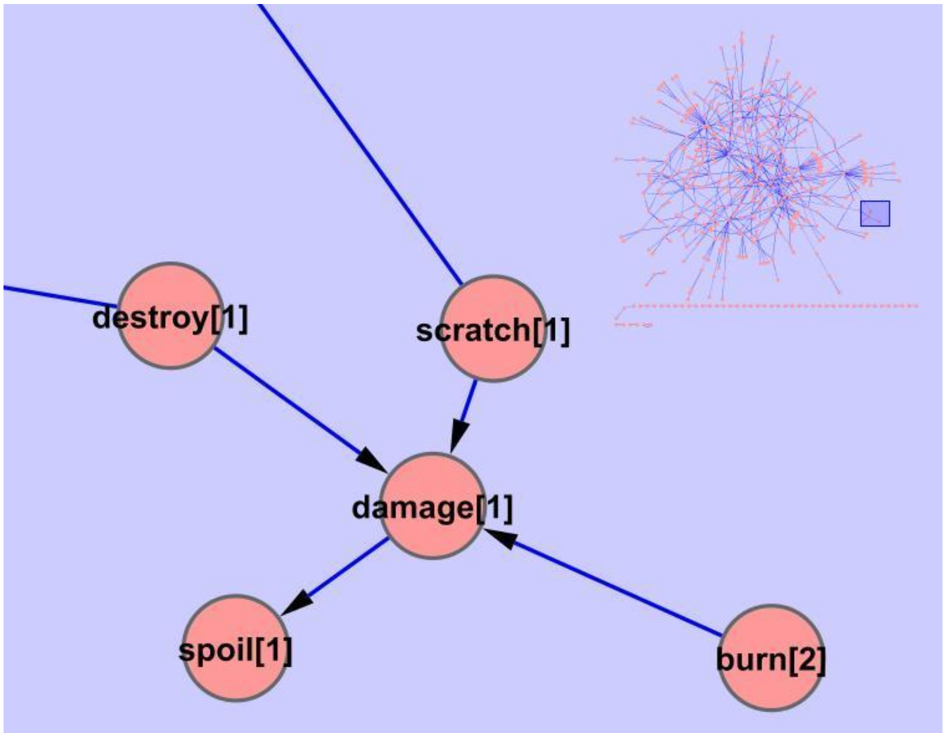

Figure 4. Example of connection between nodes in verb graph

Also, a very simple loop is shown in Fig 5. The definitions of each node is:

Pretend[1]: When you pretend, you act as though something is true when it is not really.

Act[1]: If you act, you pretend to be someone else in a play, show or film.

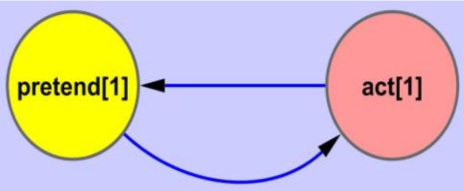

Figure 5. A simple loop in graph

Strong component or cycles in our graph analyzed and mainly 25 cycles with size 3 and 2 have found. Also nodes without out-going link (terminal) are detected. A complete list of cycles and terminals can be found in Appendices A and B.

3.2.4. Fundamental verbs

One of the objectives of this research is to not change the raw data (dictionary definition) very much and try to keep them very close to natural language. While we were expanding the definitions, we faced some verbs which if we expand it, the result would sound more complex to understand compare to the original verb or the definition is unable to transmit the verb properly. So we collect a set of verbs which according to our judgments, they are fundamental and should not be expanded. We tried to find reasons why a verb can be fundamental. We present these reasons in three classes. It is mostly the definition of a verb which causes a verb to be fundamental or not. So we made some conditions for the definition of a verb and check if the definition has these conditions. The definition of a fundamental verb can have one or more of these features:

1. A long description is needed to describe the verb.

    a. Example 1: **Put[1]:** When you put something somewhere, you move it there and leave it there.
    b. Example 2: **Change[1]:** it becomes different than what it was before [and is something new].

2. The definition is more complex to understand.

    a. Example 1: **Say[1]:** When you say something, you use your voice to make words.

3. It cannot transfer the meaning of the verb accurately and nicely.

    a. Example 1: **Think[1]:** use your mind.

These properties could help us to find fundamental verbs. Also, we used some other hints which helped us to decide if a verb can be considered as fundamental or not. For example, high frequent verb in daily English communication is a good sign for a verb to consider it as a fundamental verb. We connect fundamental verbs in our graph but we won’t expand them. The list of fundamental verb is stated in Appendix C.

3.2.5. Functional verbs

Functional verbs are a set of verbs which in most of the cases have no semantic value individually but instead, they modify or add some semantic or syntactic value to another (main) verb. Typical examples include auxiliary verbs. Additionally we made a hand-coded list of other verbs which have the same properties in our research.

Among functional verbs, we observed some which they don't act as a functional all the time. Depends on the sentence, these verbs can be functional or non-functional (a normal verb). So we split functional verbs into two sub-groups: purely functional and semi-functional.

- **Purely functional**
    - **Auxiliary verbs (Appendix D)**
    - **hand-coded functional verbs (Appendix E)**

- **Semi-functional verb**
  There exist some verbs which sometimes they act as a functional and in other cases they are normal verbs. Depends on their usage and the intentions, they can be either of them. Example:

  "Get" sometimes used instead of 'be' to form the passive.

  They're getting married later this year.

  In this case, "get" is acting like a functional.

  But "get" can acts as "receive". Example:

  I got an email from her.

  Which in this case, "got" is acting as a normal verb.

We neither connect functional verbs in our connected graph, nor expand them in expansion phase.

4. Practical aspects

4.1. How to expand

An introduction to expansion mechanism and basic concepts about it have been explained in section 3. Also in section 3.2 we became familiar with non-expandable verbs and in section 3.2.3 saw how we detect them. In this section we show how to expand the definitions and also we analyze problems which we faced during expansion phase.

429 definitions from Oxford first Dictionary (age +5) are expanded manually. As we will see, during expansion phase, some definitions needed to be split into two definitions or some new head-words were added, so the number of definitions after expansion is more than 429.

Expansion mainly contains 5 steps (some of these steps might be skipped according to the case):

1. Find the verbs in a definition.
2. Disambiguate the word senses for verbs which have been found in step 1.
3. Decide whether expand or not expand the verb. (some verbs wont needed to be expanded, see 3.2)
4. Find the corresponding definitions (body part) of the verb which have been found in step 2.
5. Replace the verbs which have found in step 1 by the definitions which have found in step 3.
6. Adjust and update the definition after replacing verbs with their definitions.

Some expansion is done without any significant problems but some of them have some difficulties. At first, some few examples of expansion which have not any specific problem are shown. Then we reveal what kind of problem one might face in expanding the definitions and what is the solution to handle them.

Expansion of the verb “crash[1]” without any significant problem:

Step 1) **Crash [1]:** When something crashes, it falls with a loud noise.	
Step2)	The verb has only 1 sense. No WSD is needed.
Step 3)	Yes, expand it
Step 4) Find the corresponding “fall”:	**Fall [1]:** it comes down suddenly.
Step 5) Replace verb with its definition :	When something crashes, it comes down suddenly with a loud noise.
Step 6)	-
Repeat steps 1..6 again until all verbs are un-expandable.	
Step 1) **Crash [1]:** When something crashes, it comes down suddenly with a loud noise. The only verb is “come”	
Step 2) Sense [1_1] is the correct sense.	**Come [1_1]:** If you come to a place, you go towards it. **Come [1_2]:** If you come to a place, you arrive there.
Step 3)	Yes, expand it.
Step 4) Find the corresponding “come” :	**Come [1_1]:** you go towards it.
Step 5) Replace verb with its definition :	When something crashes, it goes towards down suddenly with a loud noise.
Step 6)	When something crashes, it goes down suddenly with a loud noise.
Go [1] and Move [1] are in a cycle so will not expand “go” anymore.	
When something crashes, it goes towards down suddenly with a loud noise. There is no verb in definition which is expandable.	
Expansion ends.	

Expansion 1- Expansion of “crash[1]”. Verbs are underlined and their schema parts are highlighted

Here are the two versions of Crash[1], before and after expansion:

Before expansion: **Crash[1]:** When something crashes, it falls with a loud noise.

After expansion: **Crash[1]:** When something crashes, it goes down suddenly with a loud noise.

Step 1) **Mix[1]:** When you mix things, you stir or shake them until they become one thing.	
Step2) Verb “stir”	The verb has only 1 sense. No WSD is needed.
Step 3)	Yes, expand it
Step 4) Find the corresponding “stir”:	**Stir [1]:** you move it around with a spoon or a stick.
Step 5) Replace verb with its definition :	When you mix things, you move them around with a spoon or a stick or shake them until they become one thing.
Step 6)	-
Repeat steps 1..6 again until all verbs are un-expandable.	
Step 1) **Mix[1]:** When you mix things, you move them around with a spoon or a stick or shake them until they become one thing.	
Step 2) Move[2] is the correct sense.	**Move [2]:** If you move something, you take it from one place to another.
Step 3)	Yes, expand it.
Step 4) Find the corresponding “move” :	**Move[2]:** you take it from one place to another.
Step 5) Replace verb with its definition :	**Mix[1]:** When you mix things, you take them from one place to another around with a spoon or a stick or shake them until they become one thing.
Step 6)	-
Step 1) **Mix[1]:** When you mix things, you take it from one place to another around with a spoon or a stick or shake them until they become one thing.	
Step 2) Shake[1] is the correct sense.	**Shake [2]:** If you shake something, you make it shake.
Step 3)	Yes, expand it.
Step 4) Find the corresponding “shake” :	**Shake [1]:** you make it shake.
Step 5) Replace verb with its definition :	**Mix[1]:** When you mix things, you take them from one place to another around with a spoon or a stick or make them shake until they become one thing.
Step 6)	-
Step 1) **Mix[1]:** When you mix things, you take it from one place to another around with a spoon or a stick or make them shake until they become one thing.	
Step 2) Shake[1] is the correct sense.	**Shake[1]:** When a thing shakes, it moves quickly up and down or from side to side.

Step 3)	Yes, expand it.
Step 4) Find the corresponding “come” :	**Shake[1]:** it moves quickly up and down or from side to side.
Step 5) Replace verb with its definition :	When you mix things, you take them from one place to another around with a spoon or a stick or make them move quickly up and down or from side to side until they become one thing.
Step 6)	-
Step 1) **Mix[1]:** When you mix things, you take it from one place to another around with a spoon or a stick or make them shake until they become one thing.	
Step 2) Take[1_1]	**Take[1_1]:** When you take something, carry it.
Step 3)	Yes, expand it.
Step 4) Find the corresponding “take” :	**Take[1_1]:** you carry it.
Step 5) Replace verb with its definition :	When you mix things, you carry them from one place to another around with a spoon or a stick or make them move quickly up and down or from side to side until they become one thing.
Step 6)	-
Carry[1] is in a cycle with **Bring[1]** > No expansion	
Move[1] is in a cycle with **Go[1]** > No expansion	
Make is functional	
Become is functional	
No more verb to expand Expansion ends.	

Expansion 2- Expansion of “mix[1]”. Verbs are underlined and their schema parts are highlighted

Schema plays an extreme role in finding the right corresponding definition in step 2. It is a good aid not only for human (if the expansion is done manually) but for automated word sense disambiguation too. Schema is a sample of how the verb will be used in a sentence. When we have a verb in a sentence, often, by looking at how it is used in a sentence and check the schemas of it, we can decide which sense of this verb the author meant by using it. Consider the example below:

Stir[1]: When you stir a liquid or a soft mixture, you move it around with a spoon or a stick.

In this definition, a verb “move” is used. We have two different sense of “move” in our dictionary so we need to disambiguate and decide which “move” the author meant.

The two senses of “move” are defined as follow:

Move[1]: If you move, you go from one place to another.

Move[2]: If you move something, you take it from one place to another.

Now we can see that how schema part can help us to decide which “move” is the corresponding one. The verb “move” is used in definition of “stir[1]” as “you move it..” and if we compare it to the two available schemas of “move[1]” and “move[2]” easily we can see that it will mach to the second sense:

You move ≠ you move something

You move it = you move something

Schema matching will not help us to disambiguate the verbs all the time. There are some few cases which we need to disambiguate using other techniques. For example, as you might have noticed, in expansion 1, in step 2 (second iteration in expansion 1), the verb “come” has two sub-senses. These two sub-senses have the same schema parts so using schema matching technique will not help us there. In this situation, other techniques of word sense disambiguation are needed, but since we were making our models manually, for the moment we won’t go in detail regarding this problem and postponed it for our future work. More information about word sense disambiguation can be find at (Mihalcea. April 2007).

We saw in expansion 1 that some of the six main steps repeated two times and we reached to the condition which the only verb in definition is un-expandable. No manual change in definition was necessary to be done in previous example. We call this kind of expansion which no manual changes are needed or not any other problems founded, as “normal expansion”.

What are challenging in expansion mechanism are, manual changes and problems which one might see while expanding.

4.2. Problems in expansion

In expansion phase, sometimes we faced some problems which causes a bad expansion. It is not very easy to define what does a “bad expansion” mean. Perhaps the best way to describe it would be to explain what kind of problems might occur. After expanding 429 definitions, twelve different groups of problems have been detected and categorized which are shown in Table 3. The full description and at least one example of each of them are mentioned then.

Problem	Effect	Strategy
Overlapping	Duplication in content	Deletion
Cycle	Two or more options are available	Choose one of them
Repetition	Duplication in content	Deletion
Extra information imposed	Semantic restriction occurs	Manual change
Missing head-word	Couldn't expand it	-Add new head-word -Leave it as a terminal
Schema problems	Schema doesn't match	- Change the schema - Change the definition
Semantic problems	The expansion is not correct semantically	Manual change
Syntactic problems	The expansion is not correct syntactically	Manual change
Definition quality	The definition is not fluent and normal	Manual change
Missing information	Referring to objects which are not included in definition	Manual change / Leave it empty
Sub-senses	Two senses merged in one	Split or not
Extreme problem	Expansion fails	Can't do anything

Table 3 – Summary of the problems which might happen during expansion phase

- **Overlapping**

In some cases, a definition has a variable part which, when we use this definition in other definition (host), the variable takes some value from the second definition. Consider the example below:

Example 1:

Bake[1]: When you bake something, you cook it in an oven.

The only verb in this definition is "cook":

Cook[1]: to heat food *in some way* so it can be eaten.

In this definition, the variable is "in some way" which it can take different values such as "on fire", "in oil", "in boiling water", "in an oven" etc. When we expand the verb "cook", the result would look like this:

Bake[1]: When you bake something, you heat it in some way [=in an oven] so it can be eaten.

As we can see, the value for this variable is in the definition of the host verb. More precisely, “in some way” is a variable which when we put it in new definition (host), it get its value from host definition (“in some way”=”in an oven”).

This kind of overlapping doesn’t make a big problem. In fact, they are showing that in expansion mechanism, two definitions fit to each other very well. The only issue would be when we are going to make the expansion automated. In this direction, the system must be able to detect such variable and the value which they will take from the host definition and then replace it. Besides detecting the right variable and the right value for it, the only operational task would be, just replace it by its value. For example, in last case, the system needs to remove “some way” and replace it with “with fire or heat”.

Example 2:

Damage[1]: If a person or thing damages something, they spoil it *in some way*.

The variable in this definition is “in some way”. Now we try to expand the definition of “burst[1]” with the help of “damage[1]”.

Burst[1]: If someone bursts something, they damage it with fire or heat.

By expanding the verb we have:

Burst[1]: If someone burns something, they spoil it in some way [=with fire or heat].

Example 3:

Bite[1]: If you bite something, you use your teeth to damage it so that it is in pieces.

And by expanding “damage” we would have:

Bite[1]: If you bite something, you spoil it using your teeth so that it is in pieces.

In example 3, the variable “in some way” took the value “using your teeth”.

We also observed simpler similar cases which often are true for pronouns. In these cases, a pronoun is used in a definition which after we expand the definition, the reference for the pronoun emerges.

Example 4:

Chase[1]: When you chase somebody, you run after *them* and try to catch *them*.

After we expanded it in some steps, we would reach to a point which we are going to expand the verb “catch”:

Catch[1]: take and hold *something* that is moving.

In this definition, "something" is variable but when we put the definition in the host definition, it takes a value which is "them". One of the tasks for future work is to find out that "them" is referring to "somebody" in automated way.

- **Cycle**

As we have seen in previous chapter, there exist some cycles in dictionary definitions so what is consequential is, not only expand them, but we need to decide which one of head-word in cycle to use. There is a very simple way to make this decision easy and procedural. For instance, when we see verb1 in the definition which this verb makes a cycle with verb2, we can decide to keep the first verb and do not expand it. This method works well most of the time but there are some cases which there is a preference to choose one of the head-words in a cycle. This preference might come based on human judge. Although using any of the verbs in a cycle shouldn't make much difference but for us, as a human, one of them has higher weight. Generally no concrete explanation or reason can be expressed for these preferences but probably using one of them makes the definition sounds more normal compare to the other one. Below two examples are expanded which contains cycle. In first case, there is not any preference to choose between verbs and in second case, we prefer to choose one of the verbs in the cycle.

Example 1:

Clean[1]: to remove dirt and dust from something.

The only verb in the definition is "remove[1]". According to our analyzes, "remove[1]" and "take_away[1]" are in a cycle. In this case, using "remove" or "take_away" doesn't make any big difference. Both can be used alternately since they roughly (at least at children level) mean exactly the same.

Example 2:

In the definition of call[1] we have,

Call[1]: If you call someone, you speak loudly so that they will go towards you.

The verbs "go[1]" and "move[1]" are in a cycle. In this case, using "move" instead of "go" makes the definition more normal to our sense.

Evidently, it is necessary to think about this kind of problems and have a solution for them but the frequency to face such problems in our research is very low so we think, for the moment, it worth's to handle them manually since it is not very time consuming.

Also, there are some few interesting cases when we have cycle in our definitions. Sometimes, a verb is defined by another verb which the second verb itself is defined by the third and the first verb. Fig 6 shows how the connected graph of this kind of definition would look like.

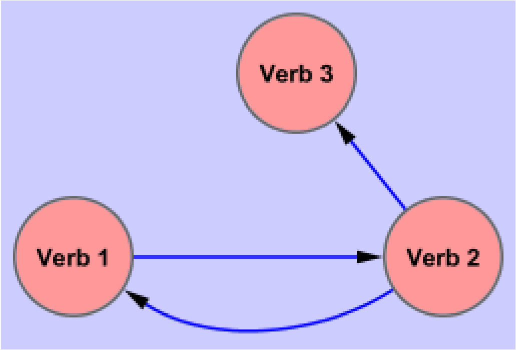

Figure 6 – special case of cycle

In this case, if we replace verb 2 by its definition, then we would have two verbs i.e. verb3 and verb1. So we can see that we are in a loop. Another option is, when we see that verb1 and verb2 are in a cycle, then we won't expand the verb2. In the second scenario, verb3 is not mentioned in expansion. A real example for such a case is divide[2] which is shown in Fig 6.

Example 3:

Divide[2] : When you divide something, you share it into equal groups.

Share[1]: If you share something, you divide it into parts and give them to other people.

If we use the first strategy and expand the "share[1]" the we would have:

Divide[2]: When you divide something, you divide it into equal groups and give them to other people.

And then we have to replace "divide" by its definition but since we already did it once before so we only use "share" instead of "divide" (they are in a cycle) and the result would be:

Divide[2]: When you divide something, you share it into equal groups and give them to other people.

If we decide not to expand the "share" then we would stick to the original definition:

Divide[2]: When you divide something, you share it into equal groups.

The main idea behind expansion is to gather as much as information we can have if the expansion doesn’t change the meaning of the original definition. So in cases similar to example 3, we prefer to use the first strategy and expand the verb.

- **Repetition**

Repetition refers to the situations which one or some parts of a definition repeat in expansion steps and it is not favorable. This problem often happens when there is an adjective or adverb. Two examples below show the problem in more detail.

Example 1:

Knock[1]: When you knock something, you hit it hard.

Hit[1]: If you hit something, you touch it hard.

By expanding the “hit” in “knock[1]”we would have:

Knock[1]: When you knock something, you touch it hard ~~hard~~.

Example 2:

Sort[1]: To arrange things in order.

Arrange[1]: If you arrange things, you put them in order.

By expanding the “arrange” in definition of “sort[1]” we would have:

Sort[1]: To you put things in order ~~in order~~.

Similar to Cycle and Overlapping, detecting and dealing with this kind of problem is not very problematic. The only necessary action to take is to delete one of the repeated words. However another way to solve this problem is to use an adverb like “very” before them (“very hard” instead of “hard hard”) but phrase like “very in order” doesn’t make sense so removing one of the repeated word sounds easier and more acceptable.

- **Schema problems**

One of the fundamental steps in expansion mechanism is, when a verb is detected in definition, we have to look for the corresponding definition of the verb and schema part is a big help to find the right definition (see 4.1). But schemas are not always easy to deal with. When we have a verb and want to check if the verb matches with the schemas, three things may happen:

1. The schema match very well and we can determine that it is the right definition to expand. Apparently, sometimes we need to disambiguate if there are similar schema for different senses.
2. The schema doesn't match at all and we know it is not the corresponding definition which we are looking for.
3. The definition which we are looking at is the corresponding and correct one but the schema doesn't match. (Another possibility which is related to case 3 is that, there are some limitations inside schema. We addressed this case separately. See next sub-parts).

In third case we need to take some actions. For instance it is necessary to make minority changes in its structure to make it suitable for expansion. Two examples show how this problem can be solved:

Example 1:

Sail[1]: To sail means to travel in a boat.

This definition has not schema part as we usually see in other definition but it is easy to change it in a way that it becomes similar to our typical schema pattern as follow:

Sail[1]: When you sail, you travel in a boat.

Also, in some cases it seems that the lexicographer assumed that it is not necessary to write the complete schema or maybe (s)he forgot.

Example 2:

Search[1]: When you search, you look very carefully for something.

In fact, this definition tries to define the verb “search for”. What makes us more certain, is the word “something” in the body part. In our view, the complete schema for this definition is:

Search[1]: When you search *for something*, you look very carefully for *something*.

Very few cases have been seen which some changes in schema part are needed to fix the problem but if there are, manual changes sounds the most practical way to handle them.

It is not always schema part which needs to be changed, sometimes a definition should be edited to fit to the schema part.

Example 3:

Argue[1]: When you argue with somebody, you talk about things *you do not agree on.*

The schema for “agree” is as follow:

Agree[1]: If *you agree with someone*, you think the same as they do.

What we can see is, the schema for “agree” is not followed in definition of “argue” but if we change the definition as follow, it would be matched exactly:

Argue[1]: When you argue with somebody, you talk about things which *you do not agree with them*.

Sometimes, we need to make smarter changes.

Example 4:

Kneel[1]: When you kneel, you change your legs shape so it is no longer straight until your knees are touching the ground.

In this example, the verb “touch” is interesting. Below two senses of “touch” are mentioned:

Touch[1]: If you touch something, you put your hand or fingers on it.
Touch[2]: If things are touching, they are so close there is no space between them.

The verb “touching” in the definition of “kneel” is referring to the second sense of touch but there is a problem with its schema part. It seems that the schema of “touching” in “kneel” definition doesn’t match the schema of “touch[2]” but we know that it is the corresponding sense.

So one solution is to change the definition as follow:

Knee[1]: When you kneel, you <u>change</u> your legs shape so it is no longer straight until *your knees and the ground are touching*.

In new definition, the schema of "touch" matches perfectly with "touch[2]". However, this example is just showing how we can change the schema part to matches it. The verb "touch" is considered as fundamental verb and we won't expand them in real work.

- **Extra information imposed**

Expansion mechanism supposes to gather information about a definition which is spread in dictionary. Collecting and including information from different parts of dictionary shouldn't lead fundamental changes in definitions. If expansion causes a definition changes in a way that the new version of definition has some addition information which was not true for original definition, then this expansion adds or removes some information and is not valid. What is very important in our research is, we trust on our raw data and try to not change it by imposing any extra information. This problem might happen if we don't think carefully about some expansions. For example, if a definition defined a fact without any assumption, adding new assumption and limiting it is an unacceptable change. While we were expanding our definitions, we found that in some cases it is possible to see such impositions:

Example 1:

Hunt[2]: When you hunt something, you <u>look</u> carefully <u>for</u> it.
Look[2]: If you look for something, you try to <u>find</u> it.

And the only definition for "find":

Find[1]: When you find something that has been lost, you <u>get</u> it back.

If we expand "look" in the definition of "hunt[2]", we would have:

Hunt[2]: When you hunt something, you try to <u>find</u> it carefully.

To expand the verb "find", we have only one option and that is "find[1]" but there would be an important question. Is it the right corresponding "find"?

Although the schema doesn't match but semantically they do. The definition of "find[1]" is somehow exactly what we are looking for with one extra restriction. The schema of "find" is limited by containing new information about the object i.e. if we "find" something, first it should have been "lost", and then we "find" it and get it back. In other word, according to this definition, "loosing something" is a condition for "finding" whereas in “hunt[2]” there is no external obligation which says something has to be lost first and then we hunt it.

The verb “find” is an example of case no 3 in previous sub-part. This kind of schemas needs special treatment. What we believe is, in fact the schema contains some extra information which is not in the structure of schema part but just added as extra information to help the user to understand *in which situation* the verb might be used. So the strategy which we use to deal with this type of schema is, we don’t consider the extra part as the schema. By this, the real schema for “find” would be:

Find[1]: Schema[When you find something] that has been lost, you get it back.

Another similar schema can be seen in the verb “fix”:

Fix[1]: If you fix something that is broken, you mend it.

Again we can see that the schema is “If you fix something” and the second part acts like an extra information which provides more information (condition) for the usage of the verb. Despite the fact that this extra information may make problem for us, it can be a very good hint for word sense disambiguation (see (Lesk. 1986)).

- **Missing head-word**

In section 2.5.1 it explained what a terminal is and why it exists. When there is a missing head-word, we need to take some actions. Mainly three options are available to deal with missing head-word:

1. Leave it as a terminal.
2. Add new head-word using another dictionary.
3. Try to change the definitions which have the missing head-word inside in a way that they don’t contain the missing head-word anymore. Thus there isn’t any definition which is referring to the missing-head-word and we don’t have to deal with it at all.

We had to prioritize these three choices somehow. What was clear for us was that we had to trust on the dictionary and the lexicographer as an expert person. So when we see a dictionary which is designed for age group +5, we deeply believe all the words which are inside the dictionary, belong to that age group and what is not inside it, according to the lexicographer, are not suitable for the mentioned age group. Of course there are some publishing limitations which lead the lexicographer to include or exclude some words but we ignored such restrictions. This assumption forced us to not choose the second strategy as much as possible and we putted it as the last option.

On the other hand, we were checking other dictionaries at different level by different publishers and found that a word can be defined in different way by not only different publishers but at different age groups with same publisher. It was a good motivation for us to try the third strategy as the first choice. It means, when we face a definition which it has a terminal verb inside, we replace the whole definition with new one which doesn't contain that specific (and of course not any other) terminal. In this direction we considered two important points: from which dictionary we should take the new definition and secondly, choose a definition which is as simple as the original definition is.

To choose a proper substitution for the original definition, we considered different dictionaries at different levels by different publishers. Table 4 shows all the dictionaries which have been checked to choose a good definition.

Publisher/Level	Level 1	Level 2	Level 3
Oxford	*	+	+
Collins	+	+	+
Chambers	-	-	+
Bounty books	+	-	-

Table 4 - Different dictionaries are considered to find a good definition[4]

The "+" sign means we had access to the dictionary and "-" means we didn't have. "*" is the original dictionary which we have worked on. Totally, 8 dictionaries have been considered in our work. To categorize them in different age group, we considered the number of head-words in each of them. For example, Level 1 dictionary contains 1500-3000 head-words. Dictionaries who have between 5000 and 10,000 are Level 2 and those with 15,000-20,000 head-words are labeled as Level 3.

Also the order of dictionaries is important too. For sake of simplicity and similarity, we start from the closest dictionary. Since the dictionary which we worked on was by Oxford at Level 1, so the first dictionary which we would look at to find a substitution definition was at same level (simplicity perspective), for example Collins Level 1. When we went through all the dictionaries at Level 1, we switched to Level 2 and start from Oxford (similarity perspective) and go on. An example for replacement of definition is "pull"

In Oxford children dictionary (the dictionary which we work on) defines it like:

Pull[1]: When you pull something, you <u>get hold</u> of it and make it <u>come</u> towards you.

"get hold (of)" is not in the original dictionary so we look for a close definition such as this one by Collins at Level 2:

Pull: When you pull something, you <u>hold</u> it firmly and <u>move</u> it towards you.

[4] Also, a level 2 dictionary by USBORNE is used in only 1 case which is not included in this table.

So we replace the original one with the new definition. Also sometimes a definition sounds simpler that another one or the original definition is expressed with the help of a symbol. Example is:

By Oxford children level:

Multiply[1]: When you multiply, you find the answer to a sum like 2 × 3 = 6.

And we found a definition by Bounty Book Level 1 easier:

Multiply: To add a number to itself a certain number of times.

Although changing the definition can be a good trick to get rid of some problems, we tried our best to keep the original definitions and not change them very much.

If none of the dictionaries could help us to solve the problem, then we keep the missing head-word as terminal or we add new head-word. Table 5 shows some information about the definitions which we had to replace, adding new head-word and terminals.

Added	Changed	Left as terminal
12 cases	63 cases	2 case

Table 5 – number of added, changed or terminal missing head-words.

Out of 429 senses, we had to add 12 new head-words and change 63 definition (either improve them or replace them with new definition from other dictionary). Only 2 head-words ("take off" and "work") were left as terminal. Added head-words and changed definitions are listed in Appendix F and G respectively. Monitoring which dictionary by which publisher is used more frequently can help us to decide our dictionaries for higher level in future work.

- **Missing information**

In expansion phase, rarely we saw definitions which some information is missed or is not available. This problem is mostly true for pronouns. Sometimes, a pronoun appears which its reference is not mentioned in the definition.

Example 1:

Buy[1]: When you buy something, you pay money to have it.

After some steps in expansion, we reach to the point which we have:

Buy[1]: When you buy something, you let *them* have money for thing, to own it.

But we don't know whom the underlined "them" is referring to. This missing reference could be a missing part in the schema but not necessarily. Because we have to consider the pronouns and their references, detecting this kind of pronoun (which appears in expansion phase) is important for our work.

- **Syntactic problem**

Replacing one word with its definition changes the structure of the sentence. So it is normal to see sometimes the expansion of a definition has syntactical problem. In our research, we didn't face any extreme cases and we believe available techniques can help us to solve syntactical problems which might happen during expansion. Nevertheless, passive sentences are the prime problems in this aspect. When a verb is used in passive form, we need to change lots of things.

Example 1: (second step in expanding the verb "earn")

Earn[1]: To get money that has been <u>given</u> to you for work that you do.

To expand the verb "give" we need to replace it and make it passive too.

Earn[1]: To get money that you have been let to have it, for work that you do.

Sometimes it is easy for human to make necessary changes and expand the definition properly but in some cases it is somehow impossible to have a correct definition at least with some simple changes. For the moment, in such cases, we left them as "extreme cases" or "failure" (see next part) but one possible solution for this problem would be, to change the passive form to the active form, and then expand the verbs:

Passive:
Earn[1]: To get money that has been <u>given</u> to you for work that you do.

Active:
Earn[1]: To get money that someone <u>gave</u> to you for work that you do.

In active form, the verb "give" is expandable easily:

Earn[1]: To get money that someone let you have it for work that you do.

- **Semantic manipulation**

Not surprisingly, in expanding the definition, some semantic problems appear. Some examples are mentioned to show how semantic problems occur and how we solved them.

Example 1:

Dive[1]: If you dive, you jump head first into water.

And the definition of “jump” is:

Jump[1]: When you jump, you go suddenly into the air with both feet off the ground.

If we expand the “jump” we would have:

Dive[1]: If you dive, you go suddenly into the air with both feet off the ground head first into water.

What is missing in this definition is a coordinate conjunction like “and” or a “then” which shows the parts “go suddenly into the air with both feet off the ground” and “head first into water” are connected to each other and happen in a sequence of time:

Dive[1]: If you dive, you go suddenly into the air with both feet off the ground and/then head first into water.

Basically, coordinate conjunctions have important value for our future work especially in linguistic inference phase.

Example 2:

Bite[1]: If you bite something, you use your teeth to **cut** *into* it.

And the definition of “cut” is:

Cut[1_1]: break something with a knife or scissors, for example.

If we expand the “cut” we would have:

Bite[1]: If you bite something, you use your teeth to break *into* something with a knife or scissors, for example.

This definition is not correct regarding semantic perspectives because of “break into”. So we need to make some changes in it. One suggested solution for it would be:

Bite[1]: If you bite something, you use your teeth to break something into pieces.

- **Definition improvement**

In some cases, the expanded definition doesn’t have any semantic or syntactic problems but it is possible to change it in a way that to improve its fluency or make it more similar to the natural language which human use to describe things. For example:

Offer[2]: If you offer to do something, you do not wait to be asked.

After expansion it would be:

Offer[2]: If you offer to do something, you do not stay to be wanted to do something.

Although this definition has not any semantic or syntactic problem but one would doubt if it is the best way to describe the verb “offer”. Judging if a definition is defined “good” or “bad” is not what we are looking for. Because the final goal of this model is to be a semantic knowledge for machine to draw some linguistic inference so maybe a definition which doesn’t sound very nice for a children or human, would be a good one for a machine to extract an answer. Nevertheless, we want to keep the definition at natural level as much as we could so finding such cases which are not defined very well according to a human is not waste of time. So we will change and improve the definitions in this case.

- **Extreme cases**

Until now, we have seen different problems and strategies to encounter them. In some cases we couldn’t do anything to expand the definition without having huge changes. What is important for us is to study the feasibility of a model which can help linguistic inferences and also prove that it is possible to build such a model automated. So having automated procedure is a crucial issue. Therefore, if we see definitions which it is not possible to expand them easily, we would count them as “un-expandable”.

Example 1:

Store[1]: If you store something, you keep it until it is needed.

And when we expand the verb “need” it would be:

Store[1]: If you store something, you keep it until you cannot do something without it.

Obviously, this expansion is incorrect and changing it and make it to something which is similar to the original definition needs lots of efforts. So this expansion failed and we label it as “extreme cases”.

Also, some few exceptions are available which are not expandable.

Example 2:

Sneeze[1]: When you sneeze, you suddenly push air out through your nose, making a loud noise.

What prevents us to expand the "push" is its definition:

Push[1]: When you push something, you use your hands to move it away from you.

According to this definition "push" is a physical action which should be done by hands which is very different from what we have in "sneeze" definition. Even at higher levels we couldn't find a definition for "push" which covers its usage in "sneeze". Despite facing such extreme cases, the result shows the frequency of them is very low and they don't bring up many concerns.

- **Splitting a sense into sub-senses**

When a verb has different senses, it means it can be defined and used for different purpose. So in lexicographers view, they are not the same and that is why they distinguished them into different senses. But sometimes, we saw some definitions which they don't cover only one sense, but several. For example:

Break[1]: If something breaks, it goes into pieces or stops working.

Surprisingly, this definition defines two different senses of "break" in one line by adding an "or" between them. What we believe is, the accurate definition for "break" should be represented in two different sub-senses as follow:

Break[1_1]: If something breaks, it goes into pieces.
Break[1_2]: If something breaks, it stops working.

To make our self sure about our decision, we refer to the other dictionaries. For example the verb "break" is defined in Oxford Level 3 as follow:

Break[1]: Make something go into pieces by dropping it or hitting it.
Break[3]: Stop working.

A very good sign for finding this type of definition is the conjunction coordinating "or". But we have to be careful that not all of the definitions which have "or" represent two or more sub-senses. "Or" can be used to show other things too. For example, it could be used to show two words are synonym or it can show different possibilities.

Example 1:

Act[1]: If you act, you pretend to be someone else in a play, show *or* film.

Apparently, this “or” is showing different possibilities. One way to understand if the “or” is acting as a connection between two different senses or different possibilities or synonyms is to check if the words around “or” are connected mutually in graph or not. When two words are connected mutually (a circle) then it is a good sign to consider them as “synonyms”.

Example 2:

Fasten[1]: If you open something, you make it no longer shut or closed.

The definitions graph supports the idea in this example nicely. Two words “shut” and “close” are connected in the graph therefore we can understand that the “or” is referring two synonyms thus we will not split it. As we saw in first example, splitting senses into sub-senses makes our mechanism more accurate and in expansion phase we can refer only to one of them but sometimes it causes problems too.

Example 3:

Crash[1_1]: When something crashes, it falls with a loud noise.
Crash[1_2]: When something crashes, it hits something else with a loud noise.

In this example, although we distinguished between two sub-senses but it is quite difficult to find out the difference between them. In the other word, one can consider them as a one definition in a way that, when something falls, it hits something else (for example ground). However, we don’t see any serious problem in these cases. If we couldn’t detect the right sub-senses we merge them and use them as it was before.

4.3. Methodology

The above mentioned problems are the most significant one hindering the expansion mechanism. One expansion can have one or several of these problems at the same time. Table 6 illustrates the frequency of each of them.

Problem	**Number of observation**
Overlapping	Many (exhaustive process)
Cycle	3
Repetition	6
Extra information imposed	9
Missing head-word	77 (Details in table 5)
Schema problems	6
Semantic problems	9
Syntactic problems	7
Definition quality	3

Missing information	2
Sub-senses	22
Extreme problem	9 (7 of them are syntactical problems)

Table 6 – Problems observed in expansion phase

The starting point and steps which we follow to expand 429 verbs are another important issue. Different methods can be used to apply expansion mechanism on data. For example we can choose a specific verb and start to expand all of the definitions which contain the verb. Another possibility is to process each verb one by one. We believe each of them has its own advantages and disadvantages. Since we start to expand them manually at first step, so we took the second option and that is expanding all the verbs one by one, sorted alphabetically. One advantage of this method is, it prevents duplicate expansions. What we mean is, assume verb1 which is already expanded in 6 steps. Later on we observe verb2 which in its definition verb1 is used. What we can do is, we use the expansion of verb1 and we don’t have to expand verb1 from the very beginning point again. So by this, we could save time and don’t have to expand a verb which has expanded before again. A positive consequence of this method is, if the expansion of verb1 has any of mentioned problems, we can solve it once and later on (most of the time) we don’t have to deal with the problems of verb1 again if it has repeated in other definitions.

The first method has some positive points too specially when the expansion mechanism is supposed to be done automated. But what we believe is the best would be the mixture of first and second methods. It means, we take one verb, expand it completely (second method) and then replace all the definitions which have the expanded verb inside (first method). Of course this method would need some adaptations when we replace one verb in all definition with its expanded one but theoretically it is one of the simplest and best expansion ways. Nevertheless, since in our study we use the second method, if we expand a verb and solve all of its expansion problems and we see the same verb again in other dictionary we won’t count the problem twice in table 6. Table 6 shows the number of times which we observed a problem in expansion of the Oxford children dictionary.

Previously we have seen that Cycles are not always problematic. We can see cycles in three classes:

A. We can choose any of the verbs in a cycle without any preferences.
B. The cycle is a bit more complex (e.g. example 3 in “cycle” sub-part of this section).
C. We have some preferences to choose one of the verbs in a cycle.

Cases “A” and “B” are not problematic neither for a human nor for a system which suppose to expand automatically. Only case “C” might be not easy for a system since it has to decide and choose one of the verbs in a cycle. Fortunately 3 cases have found which have problem similar to case “c” which is very rare.

Only 6 and 9 expansions have repetition and extra information imposed problems respectively. Problems related to schema - either it is about the schema part or there is a problem in body part - are quite infrequent too. Only 6 cases have found which have schema problem. The frequency of semantic problems is a bit more than schema problems. 9 cases have semantic problem. 6 out of 7 cases of syntactic problems are about passive and active sentences and one has another problem. Problems related to the quality of the definition or missing information in definition are very uncommon.

After adding new head-words, we started to consider splitting a sense into two or more sub-senses. Totally 22 senses have split. Out of 22, only one case split into three sub-senses and 21 of them contained two sub-senses. As the result of adding new head-words and splitting some definitions there were 463 head-words which have been analyzed. And finally, we believe in 9 cases expansion is very complex and needs high manual changes.

5. Conclusion & Future work

In conclusion, we believe we could achieve part of our general goal which supports the hypothesis directly. We found expansion as a good mechanism to gather information which is dispersed in the dictionary with the help of core words. Regarding the expansion mechanism, we were concerned about two issues: firstly, if the expansion makes big semantic changes in the original definitions and secondly if the expansion is applicable for all the definitions. Manual expansion of the verbs definitions in the children dictionary helped us to monitor all the steps and problems which may happen during the expansion. As the result, we think, expansion will not manipulate the original definition if we use some controlling mechanisms. It was shown in the section 4.2 how semantic problems might occur and discussed the solutions to encounter them. In spite of the fact that detecting such semantic changes by human is not sophisticated, we have no obvious clue if it is easy to perform them automated. Nevertheless, we are optimism regarding this problem since our experiment shows the frequency of such problems is very low.

The second issue which concerned us has the same situation. Despite different types of problems we faced during the expansion, cases which had serious problems were very uncommon. On the other hand, if a problem observed many times, a simple solution could encounter it easily. These analysis' are good reasons to believe expansion is a practical mechanism which not only can be applied on all the definitions manually, but it could be an automated procedure with high success rate.

Moreover, we deeply believe, expansion could be very unfruitful and challenging if we stick only to one source of knowledge (one dictionary). It is a mechanism that is very dependent on the data which it is applying on. It means, although the steps in expansion are defined clearly, the output of it depends on the raw data (definitions) very much. To have an expansion with high success rate, we have to make our raw data flexible. Using different dictionaries could provide such favorable flexibility in our work but we need to have some controlling too. Defining a framework of the dictionaries at different level by different publishers constructs a very flexible and enough limited structure for our work and it had a significant affect in the expansion phase.

However, difficulties in choosing the proper definition gave us this message that these selections need lots of attentions, especially when an automated procedure is supposed to perform it.

In this thesis, a small knowledge source (children dictionary) is chosen to study the feasibility of the hypothesis mentioned in the early chapter. Therefore, extending the knowledge sources to higher levels is an important future work. Since our knowledge source is dictionary definitions, increasing to higher level is clearly defined and understandable. By moving up in the age groups in dictionaries, we are extending our knowledge sources. Moreover, core words are a significant help in this direction. These words can be used to define more complex words at higher level. We believe at each level, some new core words would be added to the previous core words but it won't be too many. This is another assumption which has to be verified in the future works. A very good and similar experience in finding such core words has been done in "Word Book" of "Voice of America (VOA)" broadcasting channel. VOA has a special part entitled as Special English. "Special English[5] is VOA's method of communicating with English learners around the world in a way that is easy to understand. The vocabulary is limited to about 1,500 words" (Source: voanews.com). In their experience they could broadcast all the news since 1959 by using some limited number of words. Being able to broadcast all the news for almost 50 years is an encouraging result for us to search for core words that can be used to describe definitions at higher level.

Extending to higher levels includes much more information which makes manual processing very time and resource consuming. Constructing the models by hand is one of the main reasons why we believe previous models are not sufficient enough. Then what is really necessary is to have an automated mechanism which helps the process of building our models automatically. In this report we have shown what kind of problems we might face while expanding the definitions. Being aware of them helps us developing a system which could make expansion mechanism automated. This includes how to detect the verbs in the definitions, how to disambiguate them and how to expand them. Some of the problems such as word sense disambiguation or part of speech tagging are already studied for a long time so we can use the available techniques. On the other hand we need to develop an application which helps expansion properly. A project to develop such an application is already defined and started in AI lab at EPFL University. Although it is not fully automated for the moment, it helps human beings to expand the definition by saving time and making the expansion process easy.

The final model which we are looking for is made up of dictionary definitions and some semantic templates which we believe they exist in the definitions. These semantic templates are the key point in the linguistic inference in our project. After the knowledge model is built, a linguistic inference can be drawn by looking among semantic templates inside the model. So finding these semantic templates is quietly important. The first study about the available semantic templates in the dictionary definitions showed interesting results. Two pre-defined templates have been seen in the dictionary frequently. "Cause and Effect" and "Goal and How" are examples of them which are used to defined the head-words in dictionaries commonly. An

[5] http://media.voanews.com/documents/2009Edition_WordBook.pdf

example of “Cause-Effect” semantic template is the verb “*drown”*. The verb is defined as “**Drown:** they die under water because they cannot breathe”. In this semantic template, “die under water” is the *effect* and “cannot breathe” is the *cause*. An important task to complete our project is to explore the available semantic templates and construct our knowledge model, based on these templates.

References

A. Bharati, V. Chaitanya, and R. Sangal. 1995. *Natural Language Processing : A Paninian Perspective*. New Delhi: Prentice-Hall of India Pvt Ltd.

A. Collins, and E. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82 (6):407-428.

A.N. Kaplan, and L.K. Schubert. 2001. Measuring and improving the quality of world knowledge extracted from WordNet. In *Tech. Rep. 751 14627-0226*. New York: Univ. of Rochester.

C. Fellbaum. 1998. *WordNet : An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

D. Gelernter. 1994. *The Muse in the Machine: Computerising the Poetry of Human Thought*. New York: Free Press.

D.B. Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38 (11):33-38.

F. Suchanek, G. Kasneci, and G. Weikum,. 2007. Yago: A large ontology from wikipedia and wordnet. In *J.Web Sem*. Saarbruecken: Max-Planck-Institute for Computer Science.

G. Chierchia, B.H. Partee, and R. Turner. 1988. *Properties, Types and Meaning - Vol I + II*. Vol. 39: Springer.

G. Fliedner. 2004. Deriving FrameNet Representations: Towards Meaning-Oriented Question Answering. Paper read at International Conference on Applications of Natural Language to Information Systems (NLDB), at Salford, UK.

H. Lieberman, H. Liu, P. Singh, and B. Barry. 2004. Beating some common sense into interactive applications. *AI Magazine* 25 (4):63–76.

H. Liu, and P. Singh. 2004. ‘Commonsense reasoning in and over natural language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering systems*. Wellingtonand,New Zealand.

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

J. Kazimierczak. 1990. An approach to natural language processing in the rule-based expert system. Paper read at ACM Annual Computer Science Conference archive Proceedings of the 1990 ACM annual conference on Cooperation, at Washington, D.C.,US.

J. Markowitz, T. Ahlswede , and M. Evens. 1986. Semantically significant patterns in dictionary definitions. Paper read at Proceedings of the 24th annual meeting on Association for Computational Linguistics, at New York,US.

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute.

J.F. Allen. 2003. *Encyclopedia of Computer Science, 4th edition*. Chichester, UK: John Wiley and Sons.

J.F. Sowa 1992. Semantic Networks. In *Encyclopedia of Artificial Intelligence*, edited by S. C. Shapiro. New York: John Wiley and Sons.

L. Shi, and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. Paper read at Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics, at Mexico City, Mexico.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Paper read at Proc. of SIGDOC-86: 5th International Conference on Systems Documentation, at Toronto, Canada.

M. Rundell. 2008. *More than One Way to Skin a Cat: Why Full-Sentence Definitions have not been Universally Adopted*. Oxford: Oxford University Press.

M. Swan. 2005. *Practical English Usage*. Oxford: Oxford University Press.

P. Hanks. 1987. Definitions and Explanations. Paper read at J.M. Sinclair, at London, Collins,UK.

R. Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*. Rochester, New York,US.

R. Mihalcea. 2007. *Word Sense Disambiguation, Encyclopedia of Machine Learning*: Springer.

R. Smith. 1981. On Defining Adjectives, Part III. *Journal of the Dictionary Society of North America* 3:28-38.

R.J. Brachman, and H.J. Levesque. 2004. *Knowledge Representation and Reasoning*. San Francisco: Morgan Kaufmann Publishers.

S. Atkins, and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

T. Nasukawa, and T. Nagano. 2001. Text Analysis and Knowledge Mining System. *IBM Systems Journal* 40 (4):967–984.

W. Dolan, L. Vanderwende, and S. Richardson. 1993. Automatically deriving structured knowledge bases from on-line dictionaries. Paper read at Proceedings of theFirst Conference of the Pacific Association for Computational Linguistics, at Vancouver, Canada.

Y.K. Lee , H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. Paper read at Proceedings of the ACL-02 conference on Empirical methods in natural language processing(EMNLP), at Philadelphia,US.

Appendix A (List of Terminals)

Verb	In-degree	Out-degree
use[1]	22	0
feel[2]	10	0
try[1]	9	0
do[1_2]	7	0
have[1_1]	6	0
change[1]	5	0
make[2]	3	0
live[1]	2	0
keep[3]	2	0
spoil[1]	2	0
heat[1]	1	0
cut[1_2]	1	0
belong[1]	1	0
live[2]	1	0
own[1]	1	0
care[1_1]	1	0
grow[1]	1	0
work[?2]	1	0
matter[1]	1	0
take_off[?2]	1	0
burn[1]	1	0
freeze[2]	0	0
stick[1]	0	0
interrupt[1]	0	0
start[1]	0	0
can't[1]	0	0
wash[1]	0	0
finish[1]	0	0
go[2]	0	0
stop[3]	0	0
marry[1]	0	0
equal[1]	0	0
allow[1]	0	0
stop[1]	0	0
lead[3]	0	0
call[2]	0	0
belong[2]	0	0

match[1]	0	0
touch[2]	0	0
can[1]	0	0
set[1]	0	0
mind[1]	0	0
forgive[1]	0	0
catch[3]	0	0
protect[1]	0	0
catch[2]	0	0
mind[3]	0	0
fit[1]	0	0
balance[1]	0	0
might[1]	0	0
begin[1]	0	0
freeze[3]	0	0
end[1]	0	0
divide[1]	0	0
frown[1]	0	0

Appendix B (List of Cycles)

Size	Arc	Arc	
3	do[1_1] > spend[2]	spend[2] > use[1]	use[1] > do[1_1]
3	know[1] > find_out[1_1]	find_out[1_1] > learn[1]	learn[1] > know[1]
2	act[1] > pretend[1]	pretend[1] > act[1]	
2	appear[2] > seem[1]	seem[1] > appear[2]	
2	arrive[1] > reach[2]	reach[2] > arrive[1]	
2	beat[1] > win[1]	win[1] > beat[1]	
2	believe[1] > think[2]	think[2] > believe[1]	
2	carry[1] > take[2_2]	take[2_2] > carry[1]	
2	choose[1_1] > decide[1]	decide[1] > choose[1_1]	
2	choose[1_2] > decide[1]	decide[1] > choose[1_2]	
2	close[1] > shut[1]	shut[1] > close[1]	
2	discover[1_1] > find_out[1_2]	find_out[1_2] > discover[1_1]	
2	dress[1] > wear[1]	wear[1] > dress[1]	
2	go[1] > move[1]	move[1] > go[1]	
2	lie[1] > rest[1_1]	rest[1_1] > lie[1]	
2	sit[1] > rest[1_2]	rest[1_2] > sit[1]	
2	lift[1] > pick[2]	pick[2] > lift[1]	
2	read[1] > write[1]	write[1] > read[1]	
2	remove[1] > take_away[1]	take_away[1] > remove[1]	
2	do[1_1] > spend[2]	spend[2] > do[1_1]	
2	speak[1] > say[1]	say[1] > speak[1]	
2	spoil[2]<ruin[1]	spoil[2]<ruin[1]	
2	share[1]<divide[1]	divide[1]<share[1]	
2	get[2]<receive[1_1]	receive[1_1]<get[2]	
2	get[2]<receive[1_2]	receive[1_2]<get[2]	

Appendix C (Fundamental verbs with their definition)

Head-word	Definition
Say[1]	When you say something, you use your voice to make words.
See[1]	Know something using your eyes.
Boil[1]	When water boils, it is very hot and you can see bubbles and steam.
Like[1]	If you like somebody or something, you think they are nice.
Want[1]	When you want something, you feel that you would like it.
Hear[1]	When you hear, you take in sounds through your ears.
Put[1]	When you put something somewhere, you move it there and leave it there.
Change[1]	When things change, they become different.
Think[1]	Use your mind.
Think[2]	To believe in something.
Open[1]	If you open something, you make it no longer shut or closed.
Need[2]	If you need something, you cannot manage without it.
Eat[1]	When you drink, you swallow liquid.
Drink[1]	When you eat, you take food into your body.
Help[1]	When you help somebody, you do something useful for them.

Appendix D (Auxiliary verbs)

Auxiliary verb
Be
Can
Do
Have
May
Must
Shall
will

Appendix E (hand-coded functional verbs)

Head-word	Example
Start	as in "he starts cooking the lunch”
Begin	as in "he begins to cook the lunch"
Keep	as in "he keeps calling me every day"
Stop	as in "he stopped going to the swimming pool"
Finish	as in "he finished cleaning his room"
Make	as in "he makes me come late"
Let	as in "he lets me to take some more cake"
Help	as in “he helped them to plan their journey”
Interrupt	as in “he interrupted her while she was speaking”

Appendix F (Head-words which added)

Head-word	Publication	Level
Arrive[1]	Bounty Books	1
Heat[1]	Bounty Books	1
Spoil[2]	Bounty Books	1
Stop[3]	Collins	1
Look_after[1]	Collins	2
Find_out[1]	Collins	3
Appear[2]	Oxford	2
Look[3]	Oxford	2
Hurt[2]	Oxford	2
Think[2]	Oxford	3 (Primary Dictionary for Eastern Africa)
Take_away[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
Get[2]	Oxford	3 (Primary Dictionary for Eastern Africa)

Appendix G (All head-words which their definitions have changed)

Head-word	Publication	Level
believe[1]	Bounty books	1
blame[1]	Bounty books	1
bow[1]	Bounty books	1
break[1]	Bounty books	1
breathe[1]	Bounty books	1
care[1]	Bounty books	1
catch[1]	Bounty books	1
catch[2]	Bounty books	1
chew[1]	Bounty books	1
choose[1]	Bounty books	1
clean[1]	Bounty books	1
cook[1]	Bounty books	1
cut[1_1]	Bounty books	1
guess[1]	Chambers	3
annoy[1]	Chambers	3
hatch[1]	Chambers	3
explode[1]	Collins	1
fasten[1]	Collins	1
fill[1]	Collins	1 (new edition[6])
float[2]	Collins	1 (new edition)

[6] Collins level 1 has two editions. We used both old and new edition.

fold[1]	Collins	2
freeze[2]	Collins	2
freeze[3]	Collins	2
frown[1]	Collins	2
go[2]	Collins	3
balance[1]	Oxford	2
twist[1]	Oxford	2
land[1]	Oxford	2
lay[2]	Oxford	2
listen[1]	Oxford	2
look[1]	Oxford	2
make[1]	Oxford	2
meet[1]	Oxford	2
melt[1]	Oxford	2
multiply[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
pant[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
point[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
pour[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
protect[1]	Oxford	3 (Primary Dictionary for Eastern Africa)

pull[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
repair[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
see[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
seem[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
shine[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
sneeze[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
sort[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
spin[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
start[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
stop[2]	Oxford	3 (Primary Dictionary for Eastern Africa)
suck[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
taste[1]	Oxford	3 (Primary Dictionary for Eastern Africa)

tear[1]	Oxford	3 (Primary Dictionary for Eastern Africa)
think[1]	Oxford & manual change	3 (Primary Dictionary for Eastern Africa)
decide[1]	Some parts changed manually	
divide[1]	Some parts changed manually	
divide[2]	Some parts changed manually	
dress[1]	Some parts changed manually	
earn[1]	Some parts changed manually	
escape[1]	Some parts changed manually	
hide[1]	Some parts changed manually	
kneel[1]	Some parts changed manually	
understand[1]	Some parts changed manually	
wish[1]	USBORNE	2
yawn[1]	USBORNE	2

Appendix H (split head-words)

Head-word	Column2
answer[1_1]	When you answer, you speak when someone asks you a question.
answer[1_2]	When you answer, you speak when someone calls you.
break[1_1]	To damage something so that it is in pieces.
break[1_2]	If something breaks, it stops working.
come[1_1]	If you come to a place, you go towards it.
come[1_2]	If you come to a place, you arrive there.
crash[1_1]	When something crashes, it falls with a loud noise.
crash[1_2]	When something crashes, it hits something else with a loud noise.
curl[1_1]	If you curl up, you sit with your body bent round itself.
curl[1_2]	If you curl up, you lie with your body bent round itself.
cut[1_1]	Break something with a knife or scissors, for example.
cut[1_2]	Make a hole in something with a knife or scissors, for example.
decorate[2_1]	When people decorate a room, they make it look fresh by painting it.
decorate[2_2]	When people decorate a room, they make it look fresh by Putting paper on the walls.
discover[1_1]	When you discover something, you find out about it.
discover[1_2]	When you discover something, you see it for the first time.
do[1_1]	When you do something, you spend time on it.
do[1_2]	When you do something, you finish it.
find_out[1_1]	If you find out something, you learn something.
find_out[1_2]	If you find out something, you discover something.
have[1_1]	If you have something, it is with you.
have[1_2]	If you have something, you own it.
have[1_3]	If you have something, you feel it.
hold[1_1]	If you hold something, you have it in your hands.
hold[1_2]	If you hold something, you have it in your arms.
laugh[1_1]	When you laugh, you make sounds to show you are happy.
laugh[1_1]	When you laugh, you make sounds to show you think something is funny.
lead[1_1]	If you lead people, you go in front of them to show them where to go.
lead[1_2]	If you lead people, you go in front of them to show them what to do.
pay[1_1]	To pay means to give money for work.
pay[1_2]	To pay means to give money for things you have bought.
receive[1_1]	To receive means to get something that has been given to you.
receive[1_2]	To receive means to get something that has been sent to you.
rest[1_1]	When you rest, you lie down.

rest[1_2]	When you rest, you sit quietly.
sew[1_1]	To sew means to use a needle and cotton to join pieces of cloth together.
sew[1_2]	To sew means to use a needle and cotton to fix things on to cloth.
spell[1_1]	When you spell a word, you say letters in the right order.
spell[1_2]	When you spell a word, you write the letters in the right order.
take[2_1]	Take means to bring.
take[2_2]	Take also means to carry.
teach[1_1]	When someone teaches, they help people to understand something.
teach[1_2]	When someone teaches, they show them how to do it.
turn[1_1]	When you turn, you move round.
turn[1_2]	When you turn, you change direction.