HÁSKÓLI ÍSLANDS

Hugvísindasvið

From Talking Animal
to Talking Machine

Lexical semantic relations in WordNet

B.A. Essay in English

Tatiana Valeria Kantorovich

January 2015

University of Iceland

School of Humanities

Department of English

From Talking Animal to Talking Machine

Lexical semantic relations in WordNet

B.A. Essay in English

Tatiana Valeria Kantorovich

Kt.: 210274-5379

Supervisor: Matthew James Whelpton

January 2015

Abstract

A long time passed from the first word said by Homo sapiens to the first word said by a machine. The possibility of machines talking like men confronts us with the richness and complexity of human linguistic competence and its cognitive underpinnings. The contemporary study of linguistics attempts to explain this complexity explicitly. Part of this is lexical semantics, which studies the meanings of words and the relations between them. This essay addresses the attempt to represent one aspect of lexical semantic linguistic competence (lexical semantic relations) in a major computational resource: WordNet.

There are various kinds of relations in lexical semantics: homonymy, synonymy, antonymy, hyponymy, meronymy, and troponymy. They were used in WordNet to represent the organisation of the human lexicon. WordNet has a *synset* as a main building block. Synsets are sets of word forms that are close in meaning in context. In WordNet, nouns and verbs have taxonomic structures. The word forms are divided into domains related to a specific subject and shared features. Adjectives have a structure based on the antonymy relation where bipolar adjectives divided into clusters referring to a certain meaning. Adverbs are gathered in a single file.

Psycholinguists have often attacked the WordNet structure as a representation of human linguistic competence. However, computational linguists have found the lexical semantic database useful for machine applications and natural language processing. WordNet have been translated into many languages and combined into multilingual databases such as EuroWordNet. Each language has developed its own wordnet but they are interconnected with interlingual links.

Expand and merge approaches are used for data acquisition. The expand approach assumes bilingual translation with automatic, manual and hybrid methods to fill up gaps in data. Linguistic bias between languages can be reduced by data from sources such as Wikipedia or dictionary translation by professional interpreters. The merge approach assumes use of monolingual corpora for data acquisition.

WordNet moved from cognitive science to natural language processing. It is one of the remarkable discoveries that helped scientists to come closer to the desire to teach machines to speak.

Table of Contents

1 Introduction

A long time passed from the first word said by Homo sapiens to the first word said by a machine. The field of language studies has opened up many scientific fields. The possibility of machines talking like men confronts us with the richness and complexity of human linguistic competence and its cognitive underpinnings. The contemporary study of linguistics attempts to explain this complexity explicitly. Part of this is lexical semantics, which studies the meanings of words and the relations between them. This essay addresses the attempt to represent one aspect of lexical semantic linguistic competence (lexical semantic relations) in a major computational resource: WordNet.

Humans express themselves through language in a complex way. On the one hand, words can be spelled or pronounced alike, but vary in meaning. On the other hand, words can differ in spelling, but mean the same in context. For instance, the word *bank* can mean a financial institution in economy, or it can be a side of a river in geography. Furthermore, words can inherit meanings of other words. If something is a robin or a thrush then it is a bird. If it is a bird then it is an animal. So part of the knowledge of the meaning of the words involves a knowledge of the implied relations between them and some of these relations lead to a hierarchical organisational structure.

Complexity of linguistic competence has attracted psychologists. They were inspired by Chomsky's theory of generative grammar and studied how children learn language. The psycholinguist Miller began to build a network that would represent the human lexicon. He chose lexical semantic relations as being supported by various kinds of psychological experiment and used a taxonomic structure to describe word relations and meanings. Many scientists and students were gradually involved into the project. The net of words was named as WordNet and, finally, developed into a lexical semantic database for English. The outcome has shown that, on the one hand, lexical information can be structured, and, on the other hand, complexity of human lexical information has immense varieties of interpretations and cannot be structured explicitly. Additionally, experiments have shown agreement in the psychological reality of the relations in the WordNet model (Fellbaum (1998), in Fellbaum, pp. 89-90).

Even though the WordNet project began as an attempt to represent the cognitive organisation of the human lexicon, cognitive psychologists remained unconvinced. They denied the taxonomic structure and pointed to other experiments that called into

doubt the claim that WordNet was an effective representation of human cognitive organisation. However, computational linguists found the model useful for computer applications and natural language processing. In comparison to traditional dictionaries, which are designed for practical human use and leave all sorts of semantic information implicit, the WordNet structure is organised directly around lexical semantic information, which is represented explicitly and eliminates word-sense ambiguity and vagueness. It is essential for machines.

Its success as a computational resource is reflected in the fact that WordNet has been translated with adjustments into wordnets for various languages. Some of the wordnets have been united into multilingual databases, for instance, EuroWordNet. Each of the translated wordnets has preserved the WordNet structure, but used expand or merge approaches for data acquisition. The expand approach assumes cross-linguistic data translation. Gaps between languages can be filled up by extension from additional sources manually, automatically, or by hybrid methods. The merge approach assumes a use of monolingual resources, for instance, reuse of a printed dictionary.

Each of the methods has advantages and disadvantages. On the one hand, the expand approach allows one to build a wordnet in a short period of time, but will require adjustments that can be time-consuming and expensive. On the other hand, the merge approach will require time-consuming and expensive human labour from the begging of the project, but will presumably fit better language lexicon. Also construction of a wordnet depends on available corpora.

This thesis consists of the five sections. Section 2 discusses word meaning in lexical semantics and the main lexical semantic relations such as hyponymy, entailment, meronymy for nouns and introduction into troponymy for verbs, synonymy and antonymy, and homonymy and polysemy. Organization of lexical semantic relations in the lexical database WordNet is introduced in section 3. This section also reviews some of the assumptions relating to the psychological reality of the relations in WordNet. Section 4 compares methods of the WordNet translation into wordnets for other languages, and discuss advantages and disadvantages for data acquisition for the new wordnets. Section 5 concludes the main results of the essay.

2 Word Meaning in Lexical Semantics

For centuries scholars have attempted to find a universal definition for a word. Sapir (as cited in Saeed, 2003) states that it is psychological reality for speakers where "[t]he word is merely a form, a definitely molded entity that takes in as much or as little of the conceptual material of the whole thought as the genius of the language cares to allow" (pp. 56-7). Lexical semantics studies a word as a psychological reality. One of the ways to simulate a model of psychological reality is to represent it in a structure with relations that are taken to hold between concepts.

Miller and Fellbaum (2007) state:

… car and vehicle can be thought of as labels for two nodes in a semantic network; an arc between them represents the proposition a car is a kind of vehicle. Another kind of arc, expressing parthood, relates tire and car, expressing the fact that a tire is a part of a car, and, via inheritance, a part of all kinds of cars, such as trucks and convertibles. IS-A-KIND-OF and IS-A-PART-OF are semantic relations that holds between many pairs of word concepts, as are IS-AN-ANTONYM-OF and ENTAILS" (p. 210).

This kind of structure involves a definition for a word as a grammatically free form with at least one coherent sense that can be related to others by the major semantic relations.

2.1 Taxonomy in lexical semantics

The meaning of one word can include meanings of another. If someone says *I saw my mother just now*, you know that the speaker saw a woman. The word *mother* entails a sense of the unspoken word *woman* as part of its meaning. *Entailment* is a semantic relation where the possibility of using one word to describe something automatically implies the possibility of using another word to describe the same thing. For instance, for the verbs *to buy* and *to pay*, buying requires paying, and for *to murder* and *to kill,* murdering requires killing, and so on. It is assumed that such relationships reflect the organization of a speaker's lexical knowledge.

It is reasonable to assume shared features that link the words and, conversely, features that distinguish one concept from another. For instance, "a robin is a bird that is colourful, sings, and flies" and "birds are warm-blooded vertebrates that have beaks,

wings, and feathers, and they lay eggs." The concept *robin* shares at least three features with the concept *bird*: attributes (it is warm-blooded, vertebrate); parts (it has beak, feathers, wings); and functions (it sings, flies, lays eggs) (G. Miller (1998), in Fellbaum, pp. 29-31). The process can be continued further and the robin can be compared in its features to an animal since birds are animals, and so on.

In comparison to the robin, a chicken is also a bird that shares similar features to the robin. It is warm-blooded and vertebrate. It has a beak, feathers, and wings, and lays eggs. However, the chicken differs from the robin in the manner that it is not necessary red-breasted and, as a rule, does not fly and sing. Additionally, for humans chickens are rather associated with food. Yet despite the fact that both robins and chickens entail features of the concept *bird,* psycholinguistic evidence shows that the robin is recognized as a more typical bird then the chicken.

Since both robins and chickens share features of the more general concept *bird* and the *bird* entail features of the concept *animal,* chickens and robins also entail features of the concept *animal.* Such relations can be represented as a tree with nodes for concepts, the more general of which are closer to the root and the most specific at the leaves. These relations are named taxonomic relations and represented by a hierarchical structure, for instance, in biology. Taxonomy in lexical semantics is used to build lexical semantic relations such as hyponymy.

2.2 Hyponymy

Hyponymy is an entailment relation for nouns with inclusion and taxonomic structure. Hyponyms include the meaning of a more general word, a hypernym. For instance, a chicken and a robin are birds. Hence, the words *chicken* and *robin* are hyponyms for the hypernym *bird*. In the hierarchical structure the hypernym *bird* is a node that contains the hyponyms *chicken* and *robin*. The hypernym *bird* is a hyponym for another hypernym, for instance, *animal.* In other words, the robin and the bird are in the *is-a* relation, that is to say, *robin is-a bird.* Further, the *bird* is an *animal*, thus, the chain of the *is-a* relation can be continued as follows *robin is-a bird is-a animal*.

The taxonomic structure of hyponymy in lexical semantics is asymmetric and transitive. Asymmetry means that hypernyms and hyponyms are not interchangeable. Indeed, the sentence *All robins are birds* has a true value in comparison to the sentence *All birds are robins* that is false. According to transitivity, hyponyms *robin* and *chicken*

inherit all semantic features from hypernyms *birds* and *animal* respectively further up in the hierarchy.

There are assumptions that, on the one hand, such entailment relations could reflect speakers' lexical knowledge, but, on the other hand, concerns about typicality and membership judgments for concepts in hyponymy. For instance, Smith and Medin (as cited in Fellbaum, 1998) found in experiments that for a person "the time required to verify that a chicken is a bird is significantly longer than the time required to verify that a robin is a bird, even though *chicken* and *robin* stand in the same taxonomic relation to *bird*. The problem is not that *robin* occurs more frequently than *chicken* (it does not), but simply that robins are more typical birds than chickens are" (p. 32).

2.3 Meronymy

There is another type of entailment relation for nouns, where one noun denotes a whole and other nouns denote its parts. This kind of relation can be deduced from sentences with formulas like *An x is part of a y* or *An y has a x*. For instance, a *finger* is a part of a *hand* or a *hand* has a *finger*. Just as with hyponymy, meronymy relations can be taxonomic relations, but they are part-whole relations and less clear-cut and regular.

While hyponymy is transitive, meronymy may be transitive or intransitive. A transitive example is if a *face* has *mouth* and *mouth* has *lips*, then *face* has *lips* (on it). Thus, the noun *lips* in *lips on a face* has the same meaning as in *lips of mouth*. An intransitive example is a *hole* is a meronym of a *button*, and the *button* of *a shirt,* but the *hole* is not a meronym of the *shirt* (Saeed, 2003, p. 71). Indeed, a meaning for the word *hole* in *a hole in my button* differs from the meaning for the same word in *a hole in my shirt.*

Another aspect of meronymy is that the relation may not always be reversible. Thus, a *forest* always consists of *trees*, but *a tree* does not necessarily grow in a *forest*. While both, hyponymy and meronymy, are lexical semantic relations with hierarchical structures, hyponymy includes meaning of a general word as a node and meronymy is a whole-part relation between nouns.

2.4 Troponymy

Although much works in lexical semantics has focused on nouns, there is clear evidence that entitlement relations can be applied for verbs. In particular, from the sentence *If he limps then he walks* can be deducted that to *limp* entails to *walk*. Furthermore,

troponymy is the manner-subrelation entailment for verbs with inclusion and taxonomic structure. Additionally, troponymy resembles meronymy relation for nouns. In similarity to hyponymy for nouns, troponymy has superordinates and troponyms for verbs. In comparison to meronymy, meanings of verbs V_1 and V_2 are expressed by the formula *To V_1 is to V_2 in some particular manner*. There is more about the troponymy relations for verbs in next chapter.

2.5 Synonymy and antonymy

Often to define meaning for a word, speakers use the word's opposites. For instance, an opposite for *hot* is *cold,* for *large* is *small,* for *dead* is *alive*, for *young* is *old* and so on. Such words are called antonyms and the relation is called antonymy. Antonyms are compatible and related in meaning with their opposites.

Saeed (2003) groups antonyms into simple, gradable, reverses, converses and taxonomic sisters. Simple antonyms introduce complementary pairs or binary opposites such as *dead* and *alive*. It is impossible to be *dead* and *alive* at the same time. Gradable antonyms assume gradual development from a primal state to another and to a completely opposite final. For instance, antonyms *hot* and *cold* have stages between them like *warm, tepid* and *cool* (pp. 66-8).

Reverses distribute movement in opposite directions like *forward* is opposite to *backward.* Taxonomy has words at the same level that cannot be interchanged in a sentence and named taxonomic sisters. For instance, a week has taxonomic sisters that are the days of the week: *Sunday*, *Monday*, *Tuesday,* and so on. Another famous instance is colors that do not have opposites except for *black* and *white*. Taxonomic sisters are neither opposites nor synonyms: they are incompatible but related to each other.

Antonymy is an opposite relation to synonymy. Synonymy is a set of different words, which have the same or very similar meanings. For instance, *child*, *kid, bairn, infant*, and *baby* are words for a young human being in British English. These synonyms might be used in different situations. Some of them belong to different dialects or registers, or borrowed from another language. As Palmer states (as cited in Saeed, 2003), true or exact synonyms are very rare because they often have different distributions along a number of parameters (p. 65).

Both, antonymy and synonymy, is in a symmetric relation. That is to say if A is synonymous to B, B is synonymous to A. For instance, synonyms *people* and *folk* can

be interchanged symmetrically in a given context. For antonyms *large* and *small,* if *large* is opposite to *small*, then *small* is opposite to *large*.

2.6 Homonymy and polysemy

Homonymy and polysemy are relations that deal with unrelated and related multiple meanings respectively of a word form. Homonymy deals with words that have the same form but different meanings. Polysemy is used to describe a single word that has many meanings. Intuition and the cognitive judgements of speakers as well as the historical development of a word help to determine whether we have a single word with many meanings or different words with the same form (Saeed, 2003, p. 64). In conventional dictionaries polysemous meanings are listed together for the same entry, while homonymous meanings are given separately.

On the one hand, the noun *bank* can have unrelated meanings that are treated as homonymy. In geography, a *bank* refers to the land alongside a body of water. In economy, a *bank* is a financial institution that operates with money deposited by customers. The *bank* as an edge of a river probably originated in Old English from a Scandinavian source around c.1200. The *bank* as a financial institution has its origins from either Old Italian, or Middle French and entered English in the late 15c (Online Etymology Dictionary).

On the other hand, the financial noun *bank* can have polysemous meanings. In economy, it is an institution that provides various financial services. In gambling, it means a supply of money or things that are used as money in some games. In "something collected or stored", *bank* can be explained as an amount of something that is collected; a place where something is stored ready for use (Oxford Learner's Dictionaries).

3 Word Meaning Representation in WordNet

WordNet is a lexical semantic database for the English lexicon. The original idea for the creation of such database was based on attempts to understand how children learn new words (Miller, 2007, p. 209). In the 1980s, the psychologist George Miller has began to build a model of the lexical information stored in human memory. This simulation model is based on semantic relations and assumes the structure of WordNet as "a large network of linguistically labelled nodes" (Miller & Fellbaum, 2007, p. 210).

Each node represents a synonym set (synset) that is a set of word forms that are close in meaning and can be interchangeable in a particular context. Word forms with several meanings will appear in different synsets but for each of the synsets all the word forms share the single sense defined by that synset. Most synsets have explanatory glosses that are similar to glosses in conventional dictionaries. However, a synset is not equivalent to a dictionary entry. On the one hand, dictionaries are primarily ordered in terms of the word forms and a dictionary entry can have several different glosses for a polysemous word. On the other hand, WordNet is a database of lexical senses that are primarily organised in terms of the individual senses represented by synsets, so a synset has only a single gloss (G. Miller (1998), in Fellbaum, p. 24). Each synset represents a single sense or lexicalised concept.

For instance, Wordnet has ten different separate senses for the noun *bank.* Among the senses are *bank* as *a sloping land (especially the slope beside a body of water)* and *bank* as *a container for keeping money at home*, that would stand as homonymous lexemes in a conventional dictionary, or *bank* as *a financial institution that accepts deposits and channels the money into lending activities* and *bank* as *a building in which the business of banking transacted* that, in comparison to WordNet, would stand as a single lexeme with multiple senses (polysemous word) in the conventional dictionary.

Initially, WordNet contained only nouns. Later verbs, adjectives and adverbs were added. The synsets for each word category were entered separately, resulting in four independent networks. Main relations were hyponymy, meronymy, and antonymy. Since its first release in 1985, WordNet was updated regularly. The discussion in this chapter focuses on major semantic relations and based on the book *WordNet: an electronic lexical database* edited by Christine Fellbaum.

3.1 Semantic domains and unique beginners

The concepts that are related to a specific subject and share features are grouped together into *semantic domains*. Semantic domains are topical classifications that help to organize a large amount of data. They vary for different parts of speech and are not even in coverage. Topical classifications require *unique beginners* that can head the entire lexicon. For instance in the noun hierarchy, *entity* is a unique beginner for the domains *organism* and *object* while *abstraction* is a unique beginner for the domains *attribute, quantity, relations, time*, and so on (G. Miller (1998), in Fellbaum, p. 30). Unique beginners are chosen from the same domain. They are synsets with no superordinates. The division of the lexicon into semantic domains provides an initial, semantically based organization of the thousands of words that linked by semantic and lexical relations and, additionally, it can lead to the discovery of new relations that organize words and their meanings.

All together, WordNet has twenty-five noun unique beginners such as artefact, body, communication, food, locations, person, plant, time, and many others. Some of the beginners are close in classification to each other. In particular, eight of them concerned with nouns denoting tangible things, five related as abstractions, and three associated with psychological features. Hence, they can be grouped together. By using this approach, the number of unique beginners in WordNet was reduced to eleven (G. Miller (1998), in Fellbaum, p. 29).

Semantic domains for verbs are headed by denoting actions and events, and states. All in all, there are 14 specific semantic domains. Further, some verbs do not fit into any distinguished group. These verbs are elaborations of the concept *be* (*resemble, belong, suffice*) and comprise an extra group. This group also includes auxiliaries, control verbs (*want, fail, prevent, succeed)* and aspectual verbs (*begin*). Hence, 15 groups accommodate all verb synsets in English. It is worth mentioning, that the borders between verb domains are vague. For instance, the verb *whistle* in *The bullet whistled past him* can be classified as a verb of sound emission and as a verb of motion. Such verbs are treated as polysemous and belong to more than one semantic domain.

A unique beginner roughly corresponds to a primitive semantic component in lexical semantics. Absence of a single root verb that could head the entire verb lexicon makes difficult to find a unique beginner for the semantic domain of verbs. Fellbaum (1998, in Fellbaum) states three reasons for this fact. Firstly, verbs in English are twice

as polysemous as nouns, hence, there are too many basic senses to make it possible to single out as topmost sense from which all other descend. Secondly, it is awkward to link abstract verbs like *do* to its immediate subordinates like *communicate* and *move* because they appear further apart in the hierarchy. Thirdly, there is no psycholinguistic evidence to link verbs such as *do* and *move,* hence, there is cognitive incoherence with speakers' lexical organization (p. 72).

Therefore, it was decided to establish more meaningful unique beginners for the domains. Since not all verbs in a semantic domain can be grouped together under a unique beginner, it is possible that several trees can represent the domain. For instance, motion verbs can be presented by translational movement and movement without displacement; verbs of social interaction can be represented by subdomains like politics, work, and interpersonal relations and so on for other semantic verb domains.

3.2 Relations

There are many kinds of semantic relations in WordNet, but not all of them are equally important for different parts of speech. Evidently, synonymy is used for all of them. Since synonymy is the building block of WordNet and each synset is a sense in a context, so all word forms with more than one sense involve homonymy and are listed separately in the database. Nevertheless, WordNet assumes data search by a word form so that the different synsets are listed as polysemous. In effect, therefore, WordNet ignores the difference that the homonymy/polysemy distinction is intended to capture.

Regarding the other relations in the database, hyponymy for nouns and troponymy for verbs have taxonomic structure. Opposites are used to define antonyms and antonymy is a crucial relation for adjectives and adverbs. Furthermore, parts of speech use different semantic relations and, hence, are treated separately.

There are some disadvantages in WordNet organization. WordNet provides just paradigmatic relations among word meanings and does not provide syntagmatic relations to link synsets from different semantic domains related to a certain part of speech. Additionally to this, WordNet, also, does not contain syntagmatic relations to link concepts from different parts of speech (Fellbaum (1998), in Fellbaum, p. 9).

3.2.1 *Nouns*

Synonymy and synsets are fundamental building blocks in Wordnet as a whole and in particular for nouns. By contrast, antonymy is not a fundamental organizing relation for

nouns in WordNet. Yet, they are represented in the database, for instance, for male/female noun pairs and, in fact, oppositions must be entered separately for each of the pair.

Hyponymy is a subordination relation between synsets based on entailment. Subordination assumes a sequence of levels going from many specific terms at the lower level to a few generic terms at the top. Each top node in this hierarchy is a hypernym: a *robin* is a *bird, the bird* is an *animal,* the *animal* is an *organism*, and the *organism* is a *living entity.* Each level can be read as *is-a* or *is-a-kind-of* or *is-a-type-of* relations. It must be stressed that hyponymy is a relation between lexical concepts, not between word forms and represented in WordNet by a pointer between appropriate synsets (G. Miller (1998), in Fellbaum, p. 25). An advantage of such hierarchical construction for WordNet is to avoid circular loops that occasionally can occur in conventional dictionaries.

Meronymy is a part-whole relation between nouns. In WordNet meronymy is found primary in the `noun.body`, `noun.artifact`, and `noun.quantity` domains. Meronymy and hyponymy are comparable in their properties. They are both asymmetric, but meronymy, compared to hyponymy that is always transitive, have restrictions. Furthermore, meronymy and hyponymy are intertwined in complex ways because meronyms distinguish features that hyponyms can inherit. In the sentences: *It was a robin. The beak was injured*, for instance, "if {*beak*} and {*wing*} are meronyms of {*bird*}, and if {*robin*} is a hyponym of {*bird*}, then, by inheritance {*beak*} and {*wing*} must also be meronyms of {*robin*}" (G. Miller (1998), in Fellbaum, p. 38). Problems in establishing relations between hyponymy and meronymy arise from a tendency to attach parts high in the hierarchy. Miller (G. Miller (1998), in Fellbaum) illustrates it by the instance: "… if {*wheel*} is said to be a meronym of {*vehicle*}, then sleds cannot be vehicles. In WordNet a special synset was created for the intermediate concept, {*wheeled_vehicle*}" (p. 38). This synset is intended to facilitate the use of the database for computational applications rather than to reflect the psychological reality of lexicon.

Another problem why it is difficult to code meronomy in WordNet is related to the fact that meronymy can be intransitive and have various *is-a-part-of* relations. There are discussions in the literature about types of meronyms and considerable disagreement how to distinguish among them. However, three types of meronyms are coded in WordNet: W_m *is-a-component-part-of* W_h, W_m *is-a member-of* W_h, and W_m *is-the-stuff-*

that W_h is-made-from, where W_m stands for a meronym and W_h stands for a holonym. The most frequent among them is *is-a-component-of* and it is used as a default relation.

3.2.2 *Modifiers*

Adjective and adverbs are means of modifying or elaborating words. Adjectives modify senses of nouns, adverbs modify everything else: verbs, adjectives, other adverbs, and entire clauses. Adjective synsets in WordNet include adjectives, nouns, participles, and prepositional phrases. They are divided into two major groups descriptive and relational. Firstly, descriptive adjectives ascribe a value of an attribute to a noun. For instance, in *heavy package, heavy* is a value for attribute *weight.* Participial adjectives and quantifiers added as subclasses to descriptive adjectives. Secondly, relational adjectives derive from nouns like *electrical* in *electrical engineer* is related to *electricity.*

Antonymy is a basic semantic relation for adjectives. Antonymous adjectives introduce complementary pairs or binary opposites of an attribute. Adjectives *heavy* and *light* express opposing values for the attributes *weight.* However, it is not always possible to find a bipolar opposite for each adjective. *Ponderous* is similar in meaning to *heavy*, but it is not easy to find an antonyms for *ponderous.* The problem solved in WordNet by clustering similar descriptive adjectives by semantic similarity to a focal adjective. And further by pointing the focal adjective to its bipolar antonym. For instance, *ponderous* is similar in meaning to *heavy.* In this case, *heavy* can be a focus of a cluster and pointed to another cluster with antonyms that have as their focus *light.* Thus, an antonym for *ponderous* is *light.*

There is another group of adjectives that does not have bipolar opposites and, hence, semantic relations for them were not coded in WordNet. Adjectives in this group are gradually developed from primal state to another and to completely opposite final. For instance, antonyms *hot* and *cold* have stages between them like *warm, tepid* and *cool.* According to Miller (K. Miller (1998), in Fellbaum), "[s]ince this conceptually important relation of gradation does not play a central role in the organization of adjectives, it has not been coded in WordNet" (p. 53).

Markedness is a term in WordNet that is used to connect a value of an attribute to a descriptive adjective. *Heavy* and *light* are values for the attribute *weight.* Markedness has not been coded explicitly in WordNet, except the situation where a

noun, that names an attribute, and adjectives, that are expressing its values, are linked by a pointer (K. Miller (1998), in Fellbaum, p. 54).

The semantic organization of adverbs in WordNet is simple and straightforward. Adverbs have neither hierarchical structure like nouns or verbs, nor clusters like adjectives. All adverbs are listed in a single file `adv.all`. For lexical ("underived") adverbs, senses are displayed almost identically to coded sources where they are distinguished by numerals (K. Miller (1998), in Fellbaum, pp. 64-6). For instance, WordNet displays three senses for the adverb *then*:

- <`adv.all`>S: (adv) **then**, so, and so, (subsequently or soon afterward (often used as sentence connectors));
- <`adv.all`>S: (adv) **then** (in that case or as a consequence);
- <`adv.all`>S: (adv) **then** (at that time).

Derived adverbs are similarly coded to relational adjectives. WordNet has two senses for the adverb *plainly*:

- <`adv.all`>S: (adv) obviously, evidently, manifestly, patently, apparently, **plainly**, plain (unmistakably (`plain' is often used informally for `plainly'));
- <`adv.all`>S: (adv) **plainly**, simply (in a simple manner; without extravagance or embellishment).

These senses are similar to the two first senses from all seven for the adjective *plain*:

- <`adj.all`>S: (adj) apparent, evident, manifest, palpable, patent, **plain**, unmistakable (clearly revealed to the mind or the senses or judgment);
- <`adj.all`>S: (adj) **plain** (not elaborate or elaborated; simple).

3.2.3 *Verbs*

The different relations that organize verbs in WordNet are based on lexical entailment. Entailment for verbs resembles meronymy for nouns. Fellbaum (1998, in Fellbaum) states the resemblance in the manner that "[a]ny acceptable statement about part relations between two verbs always involves the temporal relation between the activities that the verbs denote. One activity or event is part of another activity or event only when it is part of, or a stage in, its temporal [realization]" (p. 78). Temporal inclusion in this case assumes degree of simultaneous participation of verbs in the discussion and bidirectional relations between them. For instance, verb pairs like *limp* and *walk* differ

from those like *buy* and *pay* in a way that *limp* entails *walk* and limping is properly included by walking, *buy* entails *pay* and buying is properly includes paying. In another words, someone *limps* if (during the time) he *walks,* but someone *buys* if and only if he *pays*. If in the first verb pair limping cannot occur without walking (but walking can occur without limping), in the second, buying cannot occur without paying the same as paying cannot occur without buying.

Another attempt to adopt semantic noun relations for verbs in WordNet is to apply hyponymy to verbs based on the principle of *is-kind-of* relations like in *To amble is kind of to walk.* It turns out that "the semantic distinction between two verbs is different from the features that distinguish two nouns in a [hyponymy] … relation" in the manner that lexicalization between "verb hyponyms" and their superordinates involves many kinds of semantic elaborations across different semantic domains (Fellbaum (1998), in Fellbaum, p. 79).

Compared to the features that nouns have (attributes, parts, functions), verbs have components such as *manner* and *cause*, *speed*, and *conveyance* of displacement for motion verbs, *degree of force* for verb denoting different kinds of hitting, degree of *intensity* of the action or state, and so on. The equivalent of hyponymy relations for verbs is called *troponymy;* for two verbs (V_1, V_2) the troponymy relation can be expressed by the formula *To V_1 is to V_2 in some particular manner,* e.g. *to amble is to walk in an ambling manner.* Using verbal nouns we might say that *Ambling is a kind of walking* because *Ambling is walking in some particular manner* in hyponymy. In this instance, *ambling* and *walking* are gerunds (verbal nouns) created from the related verbs. Furthermore, troponymy is a particular kind of entailment where troponym V_1 entails more general in meaning V_2 and they are temporally coextensive. In WordNet verb taxonomies based on troponymy relations tend to have no more than four levels.

Semantic opposition is also introduced for the verb lexicon in WordNet. Among relations are relations for converses, stative or change-of-state verbs, and co-troponyms (sisters) and others. Converses are antonyms that do not have a superordinate or an entailed verb: *give – take, buy – sell, teach – learn* and so on. They occur in the same semantic domain and name the same activity, but differ in mapping of thematic roles (*source* and *goal*) in the sentences where they occur. Change verbs and stative verbs are structured in similar to antonymous for adjectives with organization that is rather flat than hierarchical and without superordinates and troponyms.

Co-troponyms are semantically opposed verb pairs that different from their shared superordinate. For instance, *fail* and *succeed* entail *try*. Entailment relations for these verbs are characterized "not by temporal inclusion but by a kind of backward presupposition, where the activity denoted by the entailed verb always precedes the activity denoted by the entailing verb" (Fellbaum (1998), in Fellbaum, p. 82).

The cause relations in WordNet stand for two verb concepts, one of which causative and the other is resultative such as in the pairs *give-have, show-see.* Causative pairs are linked in WordNet by the appropriate pointer. Carter notes (as cited in Fellbaum, 1998) that "causation is a specific kind of entailment: if V_1 necessarily causes to V_2, then V_1 also entails V_2" (p. 83). Like all entailment relations, cause is unidirectional.

Fellbaum (1998, in Fellbaum) composed a classification of four kinds of entailment relations coded in WordNet. In her classification, the topmost relation is entailment. It distinguishes relations further, on the one hand, with temporal inclusion and, on the other hand, without temporal inclusion. Firstly, the relations with temporal inclusion are divided into troponymy with co-extensiveness (*limp – walk, amble – walk*) and troponymy with proper inclusion (*buy – pay, murder – kill*). Secondly, the relations without temporal inclusion are divided into backward presupposition (*fail – succeed*) and cause (*show – see*) (p. 84).

3.3 Some psycholinguistic assumptions

Cognitive scientists have attacked the hierarchical structure of WordNet. They doubted that lists of defining features could easily characterize all words. There were attempts to compare "the effect of distance in a lexical hierarchy" to "to traverse in thought". By this assumption, the time required to verify the statement *robin is a bird* is shorter than the time required to verify the statement *robin is an animal* (G. Miller (1998), in Fellbaum, pp. 31-2).

Others maintain that typicality is a more important factor than frequency and distance in a hierarchy. By this assumption, from statements *robin is a bird* and *chicken is a bird*, robins are more typical birds than chicken are, even though *chicken* and *robin* states in the same taxonomic relation to the to *bird.* In fact, "[s]tudies in which people are asked to rate typicality … show that [participants] … agree consistently about typical instances" and ratings actually not related to frequency or familiarity. Despite

unfavourable psycholinguistic judgements, the hierarchical structure seems to fit linguistic facts for nouns (G. Miller (1998), in Fellbaum, pp. 32-3).

Another aspect that concerns critics is the usefulness of WordNet for psycholinguistics studies. They state that learning by co-occurrence is more easy that by substitutability. Indeed, WordNet provides a good amount of paradigmatic associations, but there are no syntagmatic relations that would link word meanings from different semantic fields. An instance that is known as "tennis problem" illustrates the concern. As Miller (G. Miller, 1998) states:

> Suppose you wanted to learn the specialized vocabulary of tennis and asked where in WordNet to find it. The answer would be everywhere and nowhere. Tennis players in the `noun.person` file, tennis equipment is in `noun.artifact`, the tennis court is in `noun.location`, the various strokes are in `noun.act`, and so on. Nouns that co-occur in discussions of tennis are scattered around WordNet with nothing to pull them together (in Fellbaum, p. 34).

Similarly, other topics in WordNet have the same dispersed vocabularies. Certainly, this disadvantage diminishes the usefulness of WordNet and, especially, for cognitive studies.

4 Wordnets as WordNet Applications and Data Acquisition Methods

The Princeton WordNet project began as an implementation model of a lexicon that could be used by cognitive psychologists. Since then, psychological validity of the WordNet model has been called into question. Where cognitive scientists have been sceptical of the psychological reality of the Wordnet model, however, computer scientists have adopted it with enthusiasm as a useful tool in a range of natural language processing applications.

The usefulness of WordNet has attracted computational linguists to translate the database into other languages. The translation of the WordNet database is one important way to obtain a good source of lexical semantic data in a relatively short period of time. At this point it has to be taken into consideration that WordNet has been built manually from the very first word and it took years and a team of great scientists to develop the existed classification. Over time, the project has undergone changes and was developed into a database containing more than hundred thousand synsets. This chapter focuses on methods for WordNet translation into other languages, advantages and disadvantages of such methods, and introduces some successful implementations.

4.1 Expand approach for data acquisition methods

Most later projects used translations of the initial WordNet structure and various methods for data acquisition. The translation approach could be based on the assumption that most synsets in WordNet represent language-independent real-world concepts (Niemi, Linden, & Hyvärinen, 2012, p. 227). However, one language cannot be translated into another completely. The new wordnet requires new synsets for missing concepts. The synsets could be translated from English, for instance, by using bilingual dictionaries. The differences in this kind of translation would be expanded by additional data (Farreres, Rigau, & Rodffguez 1998, p. 66). Expand approach assumes that data to fill up gaps in lexicons can be collected automatically, manually, or by hybrid methods.

Data that collected automatically includes work with corpora and assumes automatic analysis of a large body of text or processing of large dictionaries. Manual construction would require human participation to create synsets. The hybrid method combines both automatic and manual data acquisition and assumes manual correction for automatically collected data. Each of the methods has advantages and disadvantages.

On the one hand, an automatic process would require less time for finding new synonym candidates. However, the automated method cannot ensure quality of the content. Thus, the raw data cannot be added directly into the database. It has to be checked manually for accuracy. At this point, a project needs human resources and becomes time-consuming. For these reasons, the costs to compile a wordnet are high. It could explain why most of the wordnets are not accessible for use without licences.

On the other hand, manual extension could be an endless process and take years to add synonyms to a wordnet. Translation could reduce the effort to build the net. Some of the gaps have to be filled up and corrected manually (Farreres et al.,1998, pp. 66-7). It requires an immense effort, time, and investment that would increase the project's cost. This approach is widely used, but it is not efficient for wordnets with large data.

4.2 Wikipedia as a source to expand data

Another aspect of bilingual translation is how to ensure quality of content for the new translated wordnet. There are methods that can help to solve related problems. For instance, professional interpreters translated synsets from English into Finish for FinnWordNet (Niemi et al., 2012, p. 227). Further, extension of the wordnet was done by the hybrid method. The source for new synonyms was Wikipedia which links articles for the same topic in different languages. Extraction of synonymous candidates was done automatically and checked manually for accuracy (Niemi et al., 2012, p. 228). The results for FinnWordNet have shown that Wikipedia can be a good source for data acquisition.

4.3 Merge approach for data acquisition

The merge approach uses monolingual data acquisition. For instance, DanNet applies a large, corpus-based monolingual dictionary of modern Danish to compile synsets (Pedersen, Nimb, Asmussen, Sørensen, Trap-Jensen, & Lorentzen, 2009, p. 270). This issue opens up discussion between "expand approach" and "merge approach" for finding synonym candidates for wordnets. Pederson (et al., 2009) states advantages and disadvantages for both methods:

> It is generally accepted that the former approach — where a wordnet is produced by translating synonym sets from Princeton WordNet to the target language — is easier, cheaper and ensures better consistency

between wordnets but on the other hand involves a genuine risk of linguistic bias. In contrast, the latter presents a more loyal picture of linguistic conceptualisation in a specific language but may for the same reason be less compatible with other wordnet structures; in addition, this strategy is more labour-intensive and thus correspondingly resource-demanding (p. 271).

Additionally, the merge approach that is applied for DanNet involves readjustments of hyponymy relations. There is no simple solution to which approach is better to use for compiling a wordnet. The decision depends on the aims of the project concerned, e.g. to compile a wordnet within a limited time frame or to ensure quality of content.

4.4 EuroWordNet

Many projects, which are based on translation of the WordNet core, have been realised for European languages. Some of them are gathered together into interconnected multilingual database named EuroWordNet. Among languages in EuroWordNet are Dutch, Italian, Spanish, German, French, one Slavic language, Czech, and one non-Indo-European, Estonian. Each language has developed its own wordnet but they are interconnected with interlingual links, *Interlingual Index* (W. Peters & I. Peters, 1998, p. 410).

Interlingual indexes link concepts by binary connection between two languages. That allows bilingual translation from one language into another without losing lexical meaning. For instance, for a word in Dutch can be found an exact equivalent in Spanish by means of coordinated interlingual links. Such indexes, also, help to ensure compatibility of the integrated wordnets. The EuroWordNet project was completed in the summer 1999 but the cooperative framework of EuroWordNet is continued through the Global WordNet Association (Global WordNet Association).

5 Conclusion

WordNet is a lexical semantic database for English that is represented in an electronic format and maps concepts onto words (Fellbaum, 2010, p. 241). In the earlier literature about WordNet, it was suggested that the model represented the psycholinguistic reality of lexicon. However, the psychological assumptions show that the design of WordNet is based on paradigmatic relations and does not accommodate direct links between word forms from different parts of speech. Hence, the database represents word knowledge rather than the world knowledge.

In fact, psycholinguists have largely ignored WordNet, but computational linguists have found it interesting. WordNet attracted them because it is organized conceptually, not alphabetically, and can be used for machine applications. Firstly, its realization resembles psychological reality more closely than a printed book (G. Miller (1998), in Fellbaum, p. 43). Secondly, scientists see WordNet as a promising tool to process language in useful ways and, perhaps even to understand it (G. Miller (1998), in Fellbaum, p. 44). Hence, WordNet contributes more to computational linguistics than cognitive theories.

WordNet represents lexical semantic information in a structured manner and eliminates word-sense ambiguity and vagueness. The database is a valuable semantic resource, for many applications in natural language processing and artificial intelligence. It is used most commonly to determine the similarity between words and in applications that require word sense disambiguation, mono- and cross- linguistic information retrieval, automatic text classification and summarization, question-answering systems, and machine translation (Fellbaum, 2010, p. 240).

Also, it is a computational resource to which the automated methods can be applied. The fact that WordNet has been translated with adjustments into wordnets for various languages reflects its success. Translations reveal that "[c]rosslinguistic wordnets show significant overlap at the top levels but diverge on the middle and lower levels, often due to language-specific lexicalization patterns" (Fellbaum, 2010, p. 241).

Designed as a project to introduce human linguistic competence, WordNet moved from cognitive science to natural language processing. It is one of the remarkable discoveries that helped scientists to come closer to the desire to teach machines to speak.

References

*EuroWordNe*t. http://www.illc.uva.nl/EuroWordNet.

Farreres, X., Rigau, G., & Rodffguez, H. (1998). Using WordNet for Building WordNets. Proceedings of the workshop: *Usage of WordNet in Natural Language Processing Systems* (pp. 65-82). Montreal, Canada.

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer.

Fellbaum, C. (Ed.). (1998)*. WordNet: and electronic lexical database.* MIT Press: Cambridge, MA.

Global WordNet Association. http://www.illc.uva.nl/EuroWordNet.

Miller, G., Fellbaum, C. (2007). WordNet then and now. *Lang Resources & Evaluation*, 41:209-214. Web.

Niemi, J., Linden, K., & Hyvärinen , M. (2012). Using a bilingual resource to add synonyms to a wordnet: Finn WordNet and Wikipedia as an instance. Proceedings of the 6th Global WordNet Conference (pp. 227-231). Matsue, Japan.

Online Etymology Dictionary. http://www.etymonline.com.

Oxford Learner's Dictionaries. http://www.oxfordlearnersdictionaries.com.

Pederson, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet: the challenge of computing a wordnet for Danish by reusing a monolingual dictionary. *Lang Resources & Evaluation*, 43:269-299. Web.

Peters, W., Peters, I. (1998). Automatic sense clustering in EuroWordNet. Proceedings of Language Resources and Evaluation Conference (pp. 409-416). Granada, Spain.

Saeed, J. (2003). *Semantics.* Oxford: Blackwell.

WordNet. A lexical database for English. http://wordnet.princeton.edu.