Proceedings ACS (2019) 321-340

I Do Believe In Word Senses

Philip C. Jackson, Jr. DR.PHIL.JACKSON@TALAMIND.COM
TalaMind LLC, PMB #363, 55 E. Long Lake Rd., Troy, MI, 48085 USA

Abstract

Understanding the meanings of words is essential for natural language understanding in cognitive systems. This paper discusses the nature and existence of word meanings, and how word meanings can be represented in artificial cognitive systems. An analysis is given of Kilgarriff's (1997) arguments that word senses do not exist, showing there is room to take exception. Kilgarriff's (2007) paper advocating Gricean semantics is compatible with a cognitive systems approach to word senses. Representation of word meanings is discussed for a research approach to a cognitive architecture for human-level artificial intelligence.

1. Introduction

The problem this paper addresses is whether and how words have meanings ('senses') that can be represented and used in artificial cognitive systems. This problem is relevant and important for advances in cognitive systems, because natural language understanding is an important ability for human cognition that remains to be fully replicated in artificial systems, and because understanding the meanings of words is essential for understanding natural language expressions in general.

2. Theoretical Issues and Approach

These pages consider the following theoretical questions:

- Do words have meanings? Do words have 'senses', i.e. frequently recurring meanings?
- Can word meanings be used in multiple situations, involving different tasks?
- Are dictionary definitions useful and adequate for describing and representing word meanings?
- How can meanings of words be represented, in a cognitive architecture for human-level AI?
- How can intentions of other agents be represented in a cognitive architecture?
- To what extent do word meanings "exist"? What is the nature of a word meaning's existence?

I will summarize what seem to be the current perspectives of cognitive semantics about the first question, and then discuss each question in some detail.

A contrary perspective was given by Kilgarriff (1997), saying that word senses do not exist. This paper finds that there is room to take exception with his arguments and conclusions. There is

a simple, yet important consequence for representation of natural language syntax and semantics: A single usage of a word may refer to multiple senses of the word.

To address the other theoretical questions listed above, the following pages discuss subsequent research by Kilgarriff, as well as the writings of Peirce, Wittgenstein, Johnson-Laird, Gärdenfors, and others, to give a general perspective on the nature and representation of word meanings and intentions, and the nature of 'intersubjective' existence. This perspective's answers to the above questions will be summarized at the end of the paper.

3. Current Perspectives of Cognitive Semantics

There is not a consensus single view in modern linguistics about precisely how word senses exist and should be represented. Researchers have developed several views about the nature of cognitive semantics, e.g. that word senses exist with a radial, prototypical nature; that words may develop new meanings over time, and old meanings may be deprecated; that word meanings are often metaphorical or metonymical and may involve mental spaces and conceptual blends (Fauconnier and Turner, 2002); that commonsense reasoning and encyclopedic knowledge may be needed for disambiguation, relative to situations in which words are used, and that the meanings of words and sentences may depend on the intentions of speakers.

This variety of views does not imply that words do not have meanings. Rather, it suggests there are a variety of ways in which words can have meanings, all of which would need to be supported by a cognitive system which understands natural language.

To the extent that scientists accept this variety of views, the issues raised by Kilgarriff in 1997 may be moot. Also, much modern work on computational linguistics is corpus-based and does not use word definitions, although a subcommunity within computational linguistics conducts research on word sense disambiguation reported in annual SemEval workshops.

However, the problems of how to represent the different ways in which words can have meanings, and how to support learning and using new word meanings, are still important for design of cognitive systems that may eventually achieve human-level AI. These considerations motivate writing this paper for a general audience of cognitive scientists.

4. Do Words Have Meanings?

4.1 Kilgarriff's 1997 Paper on Word Senses

The corpus linguist and lexicographer Adam Kilgarriff published a paper entitled "I don't believe in word senses". The title quoted a statement by Beryl T. (Sue) Atkins[1] in response to a discussion that presumed discrete and disjoint word senses at 'The Future of the Dictionary' workshop in Uriage-les-Bains, in October 1994.

Kilgarriff's paper went beyond the presumption of discrete and disjoint word senses, and claimed that word senses exist only relative to a task, and only according to the purposes of those who cluster citations in a corpus. This is in contrast to the idea that useful word senses can be

[1] Past President, European Association for Lexicography; General Editor, Collins-Robert English/French Dictionary; Lexicographical Adviser, Oxford University Press.

taken from general-purpose lexical resources, such as dictionaries. And it is in contrast to the idea that people may develop shared meanings for words without clustering citations in a corpus. Regarding word sense disambiguation (WSD) in natural language processing (NLP), Kilgarriff wrote:

> The implication for WSD is that there is no reason to expect a single set of word senses to be appropriate for different NLP applications. Different corpora, and different purposes, will lead to different senses. In particular, the sets of word senses presented in different dictionaries and thesauri have been prepared, for various purposes, for various human users: there is no reason to believe those sets are appropriate for any NLP application.

Kilgarriff concluded by giving broader statements that word senses do not exist:

> The scientific study of language should not include word senses as objects in its ontology. Where 'word senses' have a role to play in a scientific vocabulary, they are to be construed as abstractions over clusters of word usages. The nontechnical term for ontological commitment is 'belief in', as in "I (don't) believe in ghosts/God/antimatter". One leading lexicographer doesn't believe in word senses. I don't believe in word senses, either.

This is a remarkable statement, coming from a very distinguished lexicographer (who cites another very distinguished lexicographer) since the profession of lexicography is to write dictionaries that give definitions of word meanings, i.e. word senses. Yet I am not aware of previous papers directly reviewing Kilgarriff's analysis or challenging his conclusions. So, the following subsections will review his analysis.

4.2 Word Sense Disambiguation and Cognitive Linguistics

Kilgarriff summarizes previous research on word sense disambiguation in section 2 of his 1997 paper, and then presents an "antithesis": He discusses[2] work in cognitive linguistics showing how metaphors can allow words to "spawn additional meanings". He notes that "the structures underlying the distinct meanings of words are at the heart of the cognitive linguistics enterprise." He says this theoretical work on cognitive linguistics gives reasons to believe the word senses defined in dictionaries are not computationally relevant or useful to support WSD.

However, Geeraerts (2001) wrote:

> …a number of existing definitional and descriptive practices in the dictionary that are somewhat suspect from an older theoretical point of view receive a natural interpretation and legitimacy in the theoretical framework offered by Cognitive Semantics. ...there are three aspects of the Cognitive conception of lexical semantic structure that have to be discussed: the importance of prototypicality effects for lexical

[2] This paper often cites Kilgarriff's beliefs and statements in the present tense. Sadly, he passed away in 2015.

> structure, the intractability[3] of polysemy, and the [radial] structured nature of polysemy. …each of these points inspires a specific conclusion for lexicographical practice, or at least…it vindicates existing aspects of lexicographical practice.

Geeraerts discusses the relationship of dictionary definitions to cognitive categories in detail. His paper suggests a broader computational representation of word senses may be useful to support WSD and natural language understanding, based on the methods and structures of cognitive linguistics plus definitions in natural language such as those in dictionaries or other texts.

4.3 Word Senses and Ambiguity Tests

In section 3 of his 1997 paper, Kilgarriff discusses the question 'what is a word sense?' – He writes:

> To know what a word sense s_1 is, is to know which uses of the word are part of s_1 and which are not, probably because they are part of s_i where $i \neq 1$. If we are to know what word senses are, we need operational criteria for distinguishing them.

Depending on its interpretation, this definitional statement may have an implicit problem: It may presume that a use of a word can only have a single word sense, and that a word cannot be used with multiple senses at the same time.

Later in his paper, Kilgarriff gives an example illustrating this problem:

> The newspaper costs 25p and sacked all its staff.

This example uses *newspaper* with two senses at once, referring to a single copy of the newspaper, and to the business that produces newspapers. (The second sense is metonymically related to the first.) This example does not prove the concept of word sense is meaningless: It only shows that a single usage of a word can refer to two word senses.[4]

However, Kilgarriff takes a different view of this example. He writes:

> It is anomalous. We cannot place the origin of the anomaly in the lexicon unless we grant the word two lexical entries, one for a copy of the newspaper and one for the owner or corporate entity. Then the size of our lexicon will start to expand, as we list more and more of the possible kinds of referent for the word, and still it will never be complete. So the origin of the anomaly must be the interpretation process. But the anomaly seems similar to the anomaly that occurs with *bank*. In a case lying between *newspaper* and *bank*, how would we know whether the source of the anomaly was the lexicon or the interpretation process?

[3] Geeraerts wrote that his use of the term 'intractability' referred to "distinctiveness between senses of a lexical item [being] to some extent a flexible and context-based phenomenon."

[4] The sentence could of course be translated to a slightly more verbose sentence: The newspaper costs 25p and the publisher sacked all its staff.

Kilgarriff gave the following example for two usages of *bank*:

> Have you put the money in the bank?
> The rabbit climbed up the bank.

He discussed these sentences as an example of context (e.g. other words in a sentence) selecting different word senses of *bank*.

Both the *bank* and *newspaper* examples are cases where a word can have two different meanings. The *newspaper* example shows a single usage of a word simultaneously having two different word senses, while the *bank* example shows two usages each with different word senses.

For the *newspaper* example, it could be appropriate to add a sense to the dictionary, saying "newspaper" can refer to the publisher of a newspaper. Or a cognitive system could generate a metonymical interpretation of "newspaper" as "publisher of a newspaper" and dynamically use the metonym as a second sense of the word in the sentence.

Kilgarriff's concern about the size of the lexicon growing and never being complete is not by itself a valid argument against the existence of word senses. It's a fact of life for a natural language like English that the lexicon is constantly growing. The growth rate and size of the lexicon can be reduced if a system supports metonymical disambiguation for usages like the example of newspaper, and allows a single usage to refer to multiple word senses. However, one must expect that a frequently used word will tend to acquire new word senses. Indeed, Kilgarriff (1997) acknowledges this and discusses the growth of word senses later in his paper.

It's noteworthy that in discussing this example Kilgarriff (1997) appears to allow a single usage of a word (*newspaper*) to have two different senses, supporting the two clauses of the sentence. In discussing other example sentences, he appears to disallow this. For example, he discusses the sentence:

> Mary arrived with a pike and so did Agnes.

He writes that this sentence "could mean that each arrived with a carnivorous fish, or that each arrived bearing a long-handled medieval weapon, but not that the one arrived with the fish and the other with the weapon."

However, in a humorous situation the sentence could mean that one person arrived with a fish and the other with a weapon.[5] Jokes often rely on a single usage of a word having two different meanings. The fact that a situation is humorous and a sentence expresses a joke does not make its use of multiple word senses less valid. And such usages may occur in serious cases, like the *newspaper* example.

Kilgarriff compares the sentence about Mary and Agnes with two sentences having the same syntax:

> Tom raised his hand and so did Dick.
> Ellen bought some beans and so did Harry.

[5] For example, suppose Mary and Agnes each received invitations to a party for computational linguists that just said "Bring a pike."

He observes that the sentence about Tom and Dick is ambiguous in not specifying whether the right or left hand was raised, and allows interpretations in which both people raise the same (right or left) hand or in which each person raises a different (right or left) hand.

Regarding the sentence about Ellen and Harry he writes:

> The question now is,...is it possible that Ellen bought plants and Harry, food? If so, then the conclusion to be drawn from the test is that *bean* is ambiguous between the readings, and if not, then it is not.

It appears there is actually no problem: Each of the sentences allows interpretations in which the words *pike*, *hand*, and *beans* are used with either a single word sense or two different word senses. All three sentences are ambiguous in the same way.[6] Resolving the ambiguities requires more than knowledge of syntax and word senses – it requires knowledge (or least, expectations) of the specific situations (contexts) being described by the uses of the sentences.

In general, it appears to me that Kilgarriff's discussion of these examples does not support a conclusion that word senses do not exist or that useful word senses cannot be found in general-purpose dictionaries. Rather, it shows that discussion of these issues based on ambiguity test sentences could lead to incorrect conclusions, especially if one assumes that a single word usage can only refer to a single word sense. It does not seem reasonable to make this assumption as part of an argument that word senses do not exist. Nor does it seem reasonable to assume that a hearer cannot dynamically construct a metonymical interpretation of a word sense, as in the case for *newspaper*.

It is easy to make the mistake of assuming that a word usage can only refer to a single word sense. The syntax for Tala in (Jackson, 2014) showed a single word sense as the meaning of a word usage. It was only after studying Kilgarriff's (1997) paper that I realized the syntax should support multiple word senses for a word usage.[7] This will be corrected for (Jackson, in press).

4.4 A Corpus-Based Approach

In section 4 of his paper, Kilgarriff proposes a "quite different answer to the question 'what is a word sense?'" He summarizes this approach at the outset of his paper by writing:

> I then consider the lexicographers' understanding of what they are doing when they make decisions about a word's senses, and develop an alternative conception of the word sense, in which it corresponds to a cluster of citations for a word…Citations are clustered together where they exhibit similar patterning and meaning. The various

[6] Indeed, each sentence could also be interpreted to mean the same referenced object was acted upon by the verb: Mary and Agnes might have arrived together with the same physical pike (which might have been a fish or a weapon); Tom may have raised his hand and Dick might have also raised Tom's hand, or helped Tom raise his hand; Ellen and Harry might have together bought the same physical beans (which might have been a plant or food).

[7] The Tala syntax uses a *wsense* field to indicate the senses of a word usage, and a *wreferent* field to indicate the entities to which a word usage refers. Kilgarriff's *newspaper* example shows that both fields can have multiple values, for a word in a natural language sentence.

> possible relations between a word's meaning potential and its dictionary senses are catalogued and illustrated with corpus evidence.

He further defines this approach by specifying the following steps, in section 4.2 of his paper:

> For each word, the lexicographer:
>
> 1. calls up a concordance for the word;
>
> 2. divides the concordance lines into clusters, so that, as far as possible, all members of each cluster have much in common with each other, and little in common with members of other clusters;
>
> 3. for each cluster, works out what it is that makes its members belong together, reorganising clusters as necessary;
>
> 4. takes these conclusions and codes them in the highly constrained language of a dictionary definition.

I agree this approach makes sense. However, it does not directly address the issues for word senses related to cognitive linguistics which Kilgarriff noted in section 2 of his paper.

Nor does the usefulness of this approach justify a claim that it is the only valid way of identifying word senses, or that word senses in general-purpose dictionaries are not useful, or that word senses exist "only according to the purposes of whoever clusters citations in a corpus" or "only as abstractions over clusters of word usages".

To the contrary, one may claim that clusters of word usages exist in a corpus because words have meanings which make them useful in creating a corpus. Some meanings may be general and applicable across different domains, while others may be specialized to domains.

Thus, Gärdenfors (2017) notes that children can learn word meanings without analyzing a corpus:

> Children learn a language without effort and completely voluntarily. They learn new words miraculously fast. …A single example of how a word is used is often sufficient for learning its meaning. No other form of learning is so obvious or so efficient.

4.5 How Words Acquire New Meanings

Section 5 of Kilgarriff's paper discusses how words can acquire different meanings. The word *handbag* is used as an example. In addition to its most frequent usage ("a small bag, used by women to carry money and personal things"), he describes how it acquired a word sense connoting a weapon, and a word sense connoting a music genre. This is an interesting discussion, but it does not show that word senses do not exist, or justify a claim that word senses in general-purpose dictionaries are not useful. It does show that words can easily acquire new meanings.

4.6 Word Senses and Task Domains

In section 6 of his paper, Kilgarriff draws an implication from his previous discussions that:

> There is no reason to expect the same set of word senses to be relevant for different tasks.

This appears to be an overstatement. While many words have different meanings in different domains and for different tasks, some words can help people use general knowledge, or knowledge about other tasks, to learn a new task. People can learn about a new domain using metaphorical expressions with words from another domain, or using words that have similar meanings in different domains. Thus, some word senses can be relevant for different tasks.

4.7 Summary of Kilgarriff's 1997 Discussion

Section 7 of Kilgarriff's 1997 paper summarizes his discussions in previous sections. For the reasons I've given above, there is room to take exception with his conclusion that "The scientific study of language should not include word senses as objects in its ontology". One may question a belief that useful word senses are not given in general-purpose lexical resources, such as dictionaries, and question a belief that word senses exist only relative to a task, and only according to the purposes of those who cluster citations in a corpus.

One may contend that word senses exist according to the purposes of everyone who learns and uses a natural language, and for everyone who explains or defines a word for someone else.

Initial answers to the first three questions for this present paper can be summarized as follows:

- *Do words have meanings?* Yes. A word can have multiple different meanings or senses.
- *Can word meanings be used in multiple situations, involving different tasks?* Yes. In general, words are especially useful when they have meanings that support usage in different tasks, although many words have meanings that are specialized to tasks and domains.
- *Are dictionary definitions useful and adequate for describing and representing word meanings?* They can be useful and adequate for describing and representing some word meanings, especially since people can dynamically construct metonymical interpretations. However, some word meanings may be best represented nonlinguistically, e.g. associatively or as regions in a feature-space.

To address the other questions listed at the start of this paper, the following sections will give more discussion about the nature of meanings, not limited to dictionary definitions.

5. The Nature of Meaning

If we grant that words can have meanings, then some further discussion about the nature of meaning, and different kinds of meanings, may help address questions about representations of meanings in cognitive architectures. This discussion will lead to other topics such as the representation of intentions in cognitive architectures, and the nature of a word meaning's existence. To introduce these topics, it is helpful to begin with a discussion of Kilgarriff's subsequent research.

5.1 Kilgarriff's 2007 Paper on Word Senses

After his 1997 paper, Kilgarriff continued to support research on word sense disambiguation (Evans *et al*., 2016). In a 2007 paper, he clarified his views. He did not retract his 1997 paper,[8] but he said more about the nature of word senses, advocating an approach consistent with current perspectives on cognitive semantics, such as those summarized in Section 3 above. He contrasted two different philosophical approaches to the nature of meaning: A "Fregean" approach based on reifying the meanings of natural language sentences as truth values, and a "Gricean" approach which he summarized as saying:

> The meaning of sentences can be reduced to a speaker's intention to induce a belief in the hearer by means of their recognition of that intention.[9]

Kilgarriff (2007) argued convincingly that the Fregean approach to meaning is not workable in general, and that the Gricean approach to meaning is more useful and correct philosophically. He cited Strawson's (1970) essay contrasting the two schools of thought as those including Grice, Austin, and the later Wittgenstein, versus Frege, Chomsky, and the earlier Wittgenstein. Without endorsing Strawson's characterization of Chomsky's views (which appear to have not been static) this seems to be a fair summary of two approaches to semantics.

Kilgarriff (2007) did not mention Peirce, whose much earlier discussion of semantics most naturally belongs in the same camp as Grice and the later Wittgenstein. A brief discussion of Peirce and Wittgenstein will set the stage for discussing how word senses exist and may be robustly represented in artificial cognitive systems.

5.2 Peirce and Wittgenstein's Theories of Understanding and Meaning

Besides understanding natural language, Peirce also considered the understanding of phenomena in general, e.g. developing and using explanations of how (by what cause) and why (for what purpose) something happens or is done. Peirce discussed language as a system of *signs*, where a sign is something that can stand for (represent) something else.[10]

Peirce described a general process by which signs are understood. He called an initial sign (thing to be understood) a *representamen*. It is typically something external in the environment. It may be a symbol printed on paper (such as a Chinese symbol for "lamp", 灯); or smoke perceived at a distance; or, to use Atkin's (2010) example, a molehill in one's lawn; or a natural language utterance (such as "a log is in the fireplace"); or anything else perceived in the environment.

The process of understanding the representamen leads the mind to conclude that it stands for (or represents, or suggests the existence of) something, called the *object*. The object of the Chinese symbol might be a real lamp, the object of the smoke might be a fire that produces it, the

[8] Kilgarriff (2007) referenced his 1997 paper as effectively supporting Gricean semantics, and cited the 1997 paper's discussion of the "sufficiently frequent and insufficiently predictable" rule for lexicalizing word senses.

[9] Kilgarriff cites (Crane, 1995: 541-42) for this summary of Gricean semantics.

[10] This section is adapted from (Jackson, 2014, §2.2.2).

object suggested by the molehill could be a mole that created it, the object of the natural language utterance could be a log in a fireplace, etc.

From Peirce's perspective, the process of understanding a sign or representamen involves developing an explanation for the meaning or cause of the sign. Peirce used the term 'abduction' to refer to reasoning that develops explanations: If one observes something surprising, B, then one considers what fact A might naturally cause or explain B, and one concludes it is reasonable to think A might be true (Peirce, CP 5.189).

So, understanding involves developing explanations for what is observed. This applies both to understanding natural language and to understanding in general for human intelligence (cf. Hobbs et al., 1993; Bunt & Black, 2000).

According to Peirce, the mind does not go directly from the representamen to the object in developing an explanation for what is observed. The mind internally creates another sign, called the *interpretant*, which it uses to refer to the object. Within the mind, the interpretant stands for, or represents, the external object that is the represented by the first sign, the representamen (Peirce, CP 2.228).

We do not have to know precisely how internal signs (interpretants) are expressed in the brain to believe some pattern of physical information must exist in the brain constituting an internal sign, providing a link between the external representamen and the external object.

Though Wittgenstein (1922) presented a purely logical description of the relationship between language and reality in *Tractatus Logico-Philosophicus*, he later restated much of his philosophy about language in *Philosophical Investigations*. A central focus of *Investigations* was the idea that the meaning of words depends on how they are used, and that words in general do not have a single, precisely defined meaning. As an example, Wittgenstein considered the word "game" and showed it has many different, related meanings. What matters is that people are able to use the word successfully in communication about many different things. Wittgenstein introduced the concept of a "language game" as an activity in which words are given meanings according to the roles that words perform in interactions between people.

It does not appear there is any fundamental contradiction between Wittgenstein and Peirce. Rather, what Wittgenstein emphasized was that an external representamen may stand for many different objects. From a Peircean perspective this implies that the representamen may have many different internal signs, or interpretants, corresponding to different external meanings in different situations. A Peircean understanding process needs to support disambiguation (via abductive inference) of different interpretants to understand what a usage of an external sign means in a particular context.

The same external sign may be used in different contexts to represent different kinds of objects. For example, the word "canvas" can refer to a large tarp in one context and to the surface for an oil painting in another context.

And external signs can be used for representing more than external, physical objects and events. External signs can also be used to represent things that exist within minds – Peirce's word 'interpretant' is one such external sign. Cognitive scientists (and people in general) use many different words and phrases (external signs) to refer to thoughts, ideas, emotions, concepts, etc. that exist within minds and brains.

5.3 Different Kinds of Meanings and Intentions

Taking the Gricean approach to meaning, there can be different kinds of meanings, since speakers can have different kinds of intentions in the ways they use words. Section 5.5 below will discuss how intentions can be represented in cognitive systems. Here I will just note that there are variations in how different kinds of intentions can be expressed linguistically by speakers. Some of the possible combinations include:

- *Factual, Literal*. A speaker's intent is to express a fact literally, e.g. when saying "The newspaper costs 25 cents."
- *Non-factual, Metaphorical.* A speaker intends to express an idea using metaphors, e.g. when saying "All the world's a stage, and all the men and women merely players."
- *Requests and Instructions.* A speaker may intend the hearer to provide information or perform some other action. A speaker may provide instructions for how to perform an action. Requests and instructions may be expressed literally and/or metaphorically ("Go fly a kite!").
- *Emotive.* A speaker may intend to express an emotion, or evoke an emotion in the hearer. Again, the expression may be literal or metaphorical.

From a Gricean perspective, different kinds of intentions may induce different kinds of belief in the hearer, if the word 'belief' is interpreted very generally to encompass these variations.

Metaphors can be difficult for corpus-based systems. For example, if the sentence "How will Washington weather tomorrow's heated [political] storm?" is typed into a current search engine, many links to weather predictions are displayed. Links are not displayed referring to political debates, even if the question specifies the word "political" for a non-weather sense of *storm*.

There are also cases where by convention a speaker may say something that is the opposite of what is intended, and the hearer understands this. For example, actors traditionally wish other actors good luck before a performance by saying "Break a leg!"

5.4 How Can Meanings Be Represented in a Cognitive Architecture?

We do not have to believe there is just one kind of information structure used to express all internal meanings – There could be different ways of representing meanings within a cognitive system.

It is natural to identify three levels of representation within an artificial cognitive system, which may be used to represent meanings of words and to represent thoughts and perceptions in general. We may consider these as architectural levels within an agent that can perceive its environment and act intelligently within its environment. I call them the linguistic, archetype, and associative levels. They are adapted from Gärdenfors (1995).[11]

[11] The TalaMind architecture (Jackson, 2014, §1.5) combines these three architecture levels with support for a natural language of thought, multiple levels of mental representation, and a self-extending 'intelligence kernel'. Gärdenfors (1995) discussed three levels of inductive inference, which he called the linguistic, conceptual, and subconceptual levels. I consider all three levels to be conceptual levels since concepts can be represented linguistically, or associatively via neural networks.

The archetype level is where categories are represented using methods studied in cognitive linguistics and semantics such as conceptual spaces, image schemas, radial categories, etc. (Evans & Green, 2006) For example, words for colors (like "mauve") may have meanings represented by conceptual spaces. (Gärdenfors, 2000)

The associative level supports representation of categories using neural networks. It may recognize instances of common classes in the environment (e.g. faces, people, animals, chairs, cars, etc.) and process speech and visual information to recognize words, symbols, sentences, etc. This can support recognition of categories at the archetype level.

At the linguistic level an artificial cognitive system represents information and performs inference using one or more symbolic languages. Depending on the particular cognitive system, a symbolic language may be a simple notation (e.g. n-tuples of symbols), or it could be a formal, logical language like predicate calculus, or in theory it could even be a natural language like English – an approach investigated by (Jackson, 2014), which will be discussed in the following pages. At this level, meanings of words and sentences can be represented by expressions in formal or natural languages.

This does not mean that word definitions must be exact or complete – they may only be illustrative and incomplete, yet still support an intelligent agent's understanding and use of word senses, in the same way that definitions printed in physical dictionaries can be useful for humans to understand and use words.

For example, consider the English noun "fare", which has four senses listed by WordNet 2.1:

1. menu, fare – (an agenda of things to do; "they worked rapidly down the menu of reports")
2. fare, transportation – (the sum charged for riding in a public conveyance)
3. fare – (a paying (taxi) passenger)
4. fare – (the food and drink that are regularly consumed)

At the linguistic level of a TalaMind architecture for a cognitive agent (Jackson, 2014), each of these word senses could be represented by a Tala symbolic expression representing a dependency grammar parse-tree for an English definition of the word sense. Thus in the TalaMind prototype's 'discovery of bread' simulation there is a step where one cognitive agent (Leo) says to another agent (Ben): *Can you turn grain into fare for people?* Ben's linguistic level has simplified definitions for "fare" as "fee for transportation" and "food for people", for example:

```
(fare
    (wusage noun)
      (subj-of
        (means
          (obj
            (food
              (wusage noun)
              (for
                (people
                  (wusage noun)]
```

Ben disambiguates "fare" in Leo's utterance as meaning "food for people", based on "food" being a current topic of discourse. (In the prototype, *food* has a simplified definition "food means any object that an animal can eat".) Ben also uses a construction to translate the English expression "turn X into Y" into "make X be Y". So, Ben translates Leo's question into *Can Ben make grain be food for people?* (Jackson, 2014, p.235). The simulation continues with Ben reasoning and trying experiments to make grain be edible for people.

An intelligent system may perform reasoning at the linguistic level and decide communication and actions to perform in the environment. Depending on the research approach, there may be significant integration across the linguistic, archetype, and associative levels. The following sections will continue to focus on representation and processing at the linguistic level.

5.5 How Can Intentions Be Represented in a Cognitive Architecture?

To support a Gricean approach to semantics, how could an artificial cognitive system represent its intentions and the intentions of humans, if it interacts with humans?

At the linguistic level of a cognitive architecture it is also natural to identify what may be called a '*conceptual framework*', i.e. an information architecture for managing an extensible collection of concepts expressed linguistically. A conceptual framework could support processing and retention of concepts ranging from immediate thoughts and percepts to long term memory, including concepts representing linguistic definitions of words, knowledge about domains of discourse, memories of past events, expected future contexts, hypothetical or imaginary contexts, etc. These may be implemented using symbolic representations and data structures.

In particular, a conceptual framework could support 'mental models' (Johnson-Laird, 1983) providing structural representations of situations involving other agents in the world.

Broadly, mental models are "iconic" – their structures correspond to structures of situations they represent. Beyond that, mental models may be more or less elaborate, depending on what needs to be represented – a typology includes simple relations, spatial, temporal, kinematic, and dynamic models. Mental models could support spatial-temporal reasoning, which is an important topic for research, outside the scope of this paper.

Johnson-Laird (1983, pp.426-427) notes that mental models can be "meta-linguistic," i.e. contain tokens representing linguistic expressions, and that mental models can be embedded within mental models (1983, pp.430-433).

So, an artificial cognitive system could use linguistic expressions within mental models (contexts), to represent what actors within a model may wish, think, perceive, or say, and nested contexts to represent what an actor may think or perceive other actors wish, think, perceive, or say. In this way, the system could represent the intentions of other agents and have a 'theory of mind' capability. Jackson (2014, p.234) discussed how a prototype demonstration system illustrated 'nested conceptual simulation' in which an agent could create nested mental contexts to represent its thoughts about other agents' actions or thoughts about its actions or thoughts.

5.6 A 'Natural Language of Thought' in Human-Level AI

Human thoughts, beliefs, and intentions may in general be better represented by natural language expressions than by formal logic: Formal logic is handicapped in representing ambiguities and

contradictions for the broad range of human intentions, thoughts, and beliefs (cf. Sowa, 2007; Jackson, 2014, pp.60-70). Natural languages have been developed for millennia to do this.

This is one reason in favor of an artificial cognitive system using an internal language based on the syntax and semantics of a human natural language, such as English, to represent the thoughts (as well as the statements) of humans interacting with the system. In principle, such an internal language could also be used as a *'natural language of thought'* by the cognitive system to represent and help develop its own thoughts. This is a central element of the 'TalaMind' approach toward human-level AI explored by (Jackson, 2014) and discussed in subsequent papers.

I should make clear that in general I am not embracing or advocating particular arguments about languages of thought in human beings. Thus, I am not endorsing Fodor's (1975 *et seq*) arguments, nor the more recent alternative proposed by Schneider (2011) for a language of thought developed to be compatible with cognitive science and neuroscience. Nor am I endorsing Carruthers' (1996) arguments for natural language playing a role in human cognition. Such arguments are very interesting, yet I am focusing only on the nature of a language of thought for an artificial cognitive system, and in particular for a system that could (at least, arguably in principle) achieve human-level artificial intelligence.

In this regard, an argument I do accept is Jackendoff's (1989) reasoning that some concepts must be expressed as sentences in a mental language: Since there are an effectively unlimited number of natural language sentences that humans could understand, and our brains are finite, it follows that "sentential concepts" must be represented internally within the mind as structures within a combinatorial system, or language. For sake of discussion, we may call this internal language a 'mentalese', again without embracing specific arguments by Fodor or others about the nature of mentalese in humans.

Plausibly however, the expressive capabilities of natural languages should be matched by expressive capabilities of mentalese, or else the mentalese could not be used to represent concepts expressed in natural language. The ability to express arbitrarily large sentence structures and to express sentences that refer to sentences is plausibly just as important in a mentalese as it is in English. The ability to metaphorically express ideas across arbitrary, multiple domains is plausibly just as important in a mentalese as it is in English.[12]

This is not to say mentalese would have the same limitations as spoken English, or any particular spoken natural language. In mentalese, sentences could have more complex, non-sequential, graphical structures not physically permitted in speech. (Jackson, 2014) discussed hierarchical list structures for representing English syntax, to facilitate conceptual processing in support of human-level artificial intelligence.

A natural language of thought in an AI system could have great flexibility and scope: Natural language has syntax and semantics that can support analogical reasoning, causal and purposive reasoning, and logical inference in virtually any domain. Natural language can be used to define and explain mathematical and scientific formalisms and theories, which could also be represented

[12] The apparently universal explanatory and representative semantics of natural language are more important than simply being able to express an unlimited number of concepts. Considering this extensible semantics, we may aptly say the potential scope of human-level intelligence with natural language is "the beginning of infinity" (cf. Deutsch, 2011).

and processed symbolically by AI systems. Natural language syntax and semantics could help an AI system perform metacognition, by enabling the expression of specific thoughts about other specific thoughts, specific thoughts about specific perceptions, etc. (Kralik *et al*., 2018) An AI system could represent thoughts about alternative meanings of words in sentences, and represent thoughts about different ways to interpret utterances. It could represent and reason about alternative ways to express thoughts in communicating with humans and other systems.

So for the examples discussed in previous sections, a future cognitive system could in principle internally create and process concepts represented in a natural language of thought corresponding to sentences such as:

> Perhaps Mary and Agnes arrived together with the same physical pike.
>
> It's likely that Mary and Agnes brought a fish rather than a weapon, since people normally bring food to parties.
>
> Mary is likely to interpret the invitation to bring a pike as referring to a fish, because invitations normally ask people to bring food to a party.
>
> Agnes might be confused by the invitation to bring a pike, because she likes to cook fish and also participates in historical costume parties featuring knights and damsels.
>
> The sentence *p suggests the publisher could not afford any staff because they could only sell the newspaper for 25p per copy.[13]
>
> It is likely that sentence *q refers to a political debate in Washington D.C. as a heated storm because politicians there are almost always arguing and the physical weather there is generally cool at this time of year.

And so forth. In principle, in creating such sentences an AI system using a natural language of thought could represent and reason about syntactic or semantic interpretation biases that people might have in understanding natural language, and represent that some ways of expressing thoughts are potentially difficult for people to understand.

There is not a consensus based on analysis and discussion among scientists that an AI system cannot use a natural language like English as an internal language for representation and processing. Rather, it has been an assumption by AI scientists over the decades that computers must use formal logic languages (or simpler symbolic languages) for internal representation and processing of thoughts in AI systems. It does not appear there is any valid theoretical reason why the syntax and semantics of a natural language like English cannot be used directly by an AI system as its language of thought, without translation into formal languages, to help achieve human-level AI (cf. Jackson, 2014, pp.153-174).

In proposing development of a natural language of thought called Tala, based on the syntax and semantics of English, (Jackson, 2014, p.75) did not prescribe any particular approach to structuring the Tala lexicon. The thesis noted the Tala lexicon could leverage network and inheritance work by previous researchers. The TalaMind architecture is also open to inclusion and use of a generative lexicon (Pustejovsky, 1995 *et seq*.) at the linguistic level, though there are some issues to discuss:

[13] The system could use internal pointers to enable sentences to refer unambiguously to specific other sentences, or to parts of sentences.

Jackson (2014, p.71) advocated support for non-grammatical natural language expressions, as well as grammatical expressions, because people frequently use non-grammatical language and a human-level AI needs to be able to represent and try to understand whatever people say. It sometimes seems that almost every example which may be cited as ungrammatical could happen in real-world usage, whether expressed by children or adults learning a language, by people speaking colloquially or poetically, people writing concise notes, etc.

So the syntax for Tala is non-prescriptive, open and flexible, e.g. by making parts of speech optional. The thesis used a dependency grammar to specify the syntax of Tala, yet noted (p.72) that its approach is open to other ways of representing syntax. The thesis (p.88) also advocated support of constructions (represented as 'executable concepts') to help process and translate Tala expressions – this was included in the prototype demonstration system.

In contrast, Pustejovsky and Batiukova (2019, p.85) write that a generative lexicon should "reject word combinations that cannot be salvaged (unless used in highly metaphorical contexts or in poetic speech): *paint the absence, *finish the blue, etc."

Yet, consider the following sequence of sentences:

> Please help complete this abstract painting.
> Look for a place where there is no paint.
> Paint the absence blue.
> Finish the blue with varnish.

Arguably, these sentences are understandable in sequence and are not highly metaphorical or highly poetic. So arguably, representing them and understanding them should not be prevented by a generative lexicon, if it is included in a TalaMind architecture. If the reader can understand these sentences, then so should a human-level AI.[14]

On the other hand, a generative lexicon could help expedite parsing and understanding natural language sentences received as word sequences from the environment. As Pustejovsky and Batiukova note, a generative lexicon could help a system "rescue (in principle) unacceptable word combinations by looking for the needed licensing elements inside the lexical entry." Thus, parsing and understanding Kilgarriff's newspaper sentence could benefit from a generative lexicon that identifies different qualia for *newspaper*, such as a constitutive role corresponding to a physical object that can have a purchase price, and an agentive role corresponding to an organization that creates and publishes newspapers, and has employees.[15]

5.7 What Is the Nature of a Word Meaning's Existence?

To further consider the nature of a word meaning's existence, here are some thoughts about the nature of existence in general and in relation to natural language. We may distinguish three modes of existence: objective, subjective, and intersubjective.

[14] Although a general lexicon would not have a word sense for *absence* as 'a place where there is no paint', this meaning is clear from the previous sentences in the example. For *absence*, the meaning 'a place where something is not present' is similar to 'a time when something is not present'. Both are supported by the usage "absence makes the heart grow fonder" which may refer to a time and place of absence.

[15] If I understand generative lexicon terminology correctly (cf. Pustejovsky and Batiukova, 2019, p.162).

Unless one is an extreme solipsist, some things have *objective* existence, physically in space and time, independently of our minds. For example, oceans, forests, and birds exist in objective reality. Arguably, their existence may not depend on whether we are aware of them or observe them, at least for most practical purposes – although at the quantum level, what exists and happens physically may depend on when and how observations are performed.

Although many people think of objective reality as the only 'reality', it's important to note that we do not have direct knowledge of objective reality. Instead, we have an internal, projected reality constructed from our perceptions of external reality (Jackendoff, 1983). Our perceptions are internal constructs that indirectly represent external reality, sometimes incompletely, inaccurately or paradoxically. It is only because our perceptions generally track objective reality very closely, that we normally think we directly perceive objective reality.

Some things have *subjective* existence within an individual mind, such as thoughts and feelings. An individual has a first-person awareness and experience of subjective existence, which other individuals can at most indirectly infer.

Some things have *intersubjective* existence, based on people sharing beliefs and ideas, primarily by using natural language. Examples include money, corporations, laws, governments, nations, etc. Such intersubjective entities may have limited grounding in objective reality, yet be very important to individuals and society. Harari (2015) discusses the importance of intersubjective existence throughout human history. (Harari's text hyphenates the term as 'inter-subjective'.)

Gärdenfors (2017) gives a detailed discussion of the role of intersubjectivity in how children learn word meanings. He notes that children rapidly learn word meanings in domains that are familiar, and have difficulty learning words in unfamiliar domains, even if the words in unfamiliar domains are used more often by adults. Thus a four-year-old child may more easily learn the meaning of a word like *mauve* or *chartreuse* than a monetary term like *mortgage* or *inflation*:

> Adding new color terms is just a matter of learning the mapping between the new words and the color space. For example, chartreuse is a kind of yellowish green, and mauve is a pale violet. On the other hand, the child is normally not acquainted with the domain of economic transactions. To the child, money means concrete things – coins and bills – that one can exchange for other things. Abstract monetary concepts are not within a child's semantic reach. Grasping a new domain is a cognitively much more difficult step than adding new terms to an already established one.

Thus, Gärdenfors (2017) hypothesizes that semantic knowledge is organized into domains, and that learning of domains is connected to the development of intersubjectivity, which involves theory of mind – the ability to represent other people's emotions, attention, desires, intentions, belief and knowledge. Gärdenfors' research discusses the use of conceptual spaces for modeling semantics of nouns, adjectives, and verbs.

In summary, word senses exist intersubjectively via natural language: People explain the meanings of words, either linguistically or by physical demonstration, and reach intersubjective

agreement that they understand the same meanings for words. Sometimes the explanations for meanings of words are written in dictionaries.

However, both intersubjective and subjective reality are grounded in objective reality, [16] since thoughts exist as physical states or processes in human brains, though we do not know precisely how they are represented. Intersubjective reality can also be partially grounded in information stored in computers, e.g. in the data representing bank accounts, organizational structures, etc.

And intersubjective existence of word senses could also be grounded in objective reality within artificial cognitive systems, by representing word meanings with physical data structures at the linguistic, archetype, or associative levels of cognitive architectures, supporting cognitive capabilities and behaviors.

6. Conclusion

To summarize, the questions asked at the beginning of this paper have been answered as follows:

- *Do words have meanings?* Yes. A word can have multiple different meanings or senses.
- *Can word meanings be used in multiple situations, involving different tasks?* Yes. In general, words are especially useful when they have meanings that support usage in different tasks, although many word meanings are specialized to tasks and domains.
- *Are dictionary definitions useful and adequate for describing and representing word meanings?* They can be useful and adequate for describing and representing some word meanings, especially since people can dynamically construct metonymical interpretations. Other words can have meanings that are best represented nonlinguistically, e.g. associatively or as regions in a feature-space.
- *How can meanings of words be represented, in a cognitive architecture for human-level AI?* Meanings of words can be represented with linguistic definitions, or as cognitive concept structures, or associatively, at the linguistic, archetype, and associative levels of a cognitive architecture.
- *How can intentions of other agents be represented in a cognitive architecture?* An artificial cognitive system could use natural language expressions within mental models (contexts), to represent what actors within a model may wish, think, perceive, or say, and nested contexts to represent what an actor may think or perceive other actors wish, think, perceive, or say.
- *To what extent do word meanings "exist"? What is the nature of a word meaning's existence?* Word meanings exist intersubjectively, within the minds of people who use words. Such intersubjective existence could also be supported, in principle, within artificial cognitive systems, where word meanings would be represented with physical data structures at the

[16] Yet again, our perception of objective, physical reality is only an internal, mental model, a fragmentary shadow of what actually exists in reality. People reach intersubjective agreement that they have the same perception of reality.

linguistic, archetype, or associative levels of cognitive architectures, supporting cognitive capabilities and behaviors.

In many ways, ideas are among the most important things that exist, since ideas can lead to actions that can have profound effects on physical reality. This is especially true of intersubjective reality. Entities that exist intersubjectively, such as word senses, money, corporations, governments, and scientific theories, can have profound effects in shaping objective reality. Natural language is a key ability we have for creating intersubjective reality.

References

Bunt, H. C. & Black, W. J. (Eds.). (2000). *Abduction, belief and context in dialogue: Studies in computational pragmatics*. Amsterdam: John Benjamins.

Carruthers, P. (1996). *Language, thought and consciousness – An essay in philosophical psychology*. Cambridge, UK: Cambridge University Press.

Crane, T. (1995). Meaning. In T. Honderich (Ed.), *The Oxford companion to philosophy*, 541–42. Oxford, UK: Oxford University Press.

Deutsch, D. (2011). *The beginning of infinity - Explanations that transform the world*. New York, NY: Viking Penguin.

Evans, R., Gelbukhy, A., Grefenstettez, G., Hanks, P., Jakubícek, M., McCarthy, D., Palmer, M., Pedersen, T., Rundell, M., Rychlý, P., Sharoff, S., & Tugwell, D. (2016). Adam Kilgarriff's legacy to computational linguistics and beyond. *CI-CLing 2016: Computational linguistics and intelligent text processing*, 3-25. New York, NY: Springer.

Evans, V. & M. Green (2006). *Cognitive linguistics – An introduction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge, UK: Cambridge University Press.

Fauconnier, G. & Turner, M. (2002). *The way we think – Conceptual blending and the mind's hidden complexities*. New York, NY: Basic Books.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (2008). *LOT2 – The language of thought revisited*. Oxford, UK: Oxford University Press.

Gärdenfors, P. (1995). Three levels of inductive inference. *Studies in Logic and the Foundations of Mathematics*, 134, 427-449. Amsterdam: Elsevier.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2017). Semantic knowledge, domains of meaning and conceptual spaces. In P. Meusburger, B. Werlen, & L. Saursana (Eds.), *Knowledge and action*, 203-219. New York, NY: Springer.

Geeraerts. D. (2001). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, 17, 6-21.

Harari, Y. N. (2015). *Sapiens: A brief history of humankind*. New York, NY: HarperCollins.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63 (1-2), 69-142.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. (1989). What is a concept, that a person may grasp it? *Mind & Language*, 4, 1-2, 68-102, Also in Jackendoff, R. (1992) *Languages of the mind*. Cambridge, MA: MIT Press.

Jackson, P. C. (2014). *Toward human-level artificial intelligence – Representation and computation of meaning in natural language*. Doctoral thesis, Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands.

Jackson, P. C. (in press). *Toward human-level artificial intelligence – Representation and computation of meaning in natural language*. Mineola, NY: Dover Publications.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Kilgarriff, A. (1997). "I don't believe in word senses". *Computers and the Humanities*, 31, 91-113.

Kilgarriff, A. (2007). Word senses. In Agirre, E., & Edmonds, P. (Eds.), *Word sense disambiguation – Algorithms and applications*, 29-46. New York, NY: Springer.

Kralik, J. D., Lee, J., Rosenbloom, P. S., Jackson, P. C., Epstein, S. L, Romero, O. J., Sanz, R., Larue, O., Schmidtke, H., Lee, S. W., & McGreggor, K. (2018). Metacognition for a Common Model of Cognition. *Procedia Computer Science*, 145, 730-739.

Peirce, C. S. (CP) *Collected papers of C. S. Peirce.* Edited by C. Hartshorne, P. Weiss, & A. Burks. Eight volumes published from 1931-1958. Cambridge, MA: Harvard University Press.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, J. and Batiukova, O. (2019). *The lexicon*. Cambridge, UK: Cambridge University Press.

Schneider, S. (2011). *The language of thought – A new philosophical direction*. Cambridge, MA: MIT Press.

Sowa, J. F. (2007). Fads and fallacies about logic. *IEEE Intelligent Systems*, March 2007, 22:2, 84-87.

Strawson, P. F. (1970). *Meaning and truth: An inaugural lecture delivered before the University of Oxford on 5 November 1969.* Gloucestershire, UK: Clarendon Press.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. Translated by C. K. Ogden, with assistance from G. E. Moore, F. P. Ramsey, & L. Wittgenstein. London: Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations*. Translated by G. E. M. Anscombe. Oxford, UK: Wiley-Blackwell.