

Understanding the Unknown:

Unattested Input Processing in Natural Language

Julia M. Taylor
RiverGlass, Inc and Purdue University
USA
jtaylor1@purdue.edu

Victor Raskin
Purdue University and RiverGlass, Inc.
USA
vraskin@purdue.edu

Abstract—This paper describes the process of deriving the meaning of an unknown word within the framework of meaning based natural language processing. It uses the clues supplied by the rest of the sentence, taking into account various degrees of possibilities of what the unknown word can mean, according to the previously acquired knowledge resources. The process of finding the meaning is incremental, and thus the derived meaning can evolve as more knowledge is gathered. The paper describes the fuzzy approach to the problem within this framework with regard to the semantic and syntactic acceptability of a sentence.

Unknown words; computing with words; ontological semantics; natural language understanding; sentence acceptability

I. INTRODUCTION

This paper attempts to bring attention to the understanding of previously unseen words based on the meaning of the rest of the sentence. We assume that such trait is necessary for any robust computer system that is capable of communicating intelligently with humans. The necessity is derived from the ongoing changes in natural languages, including constant addition of new terms or senses, i.e. the new meanings of the existing words (think, for instance, of the ever propagated government generated acronyms or of teenage slang terms).

The use of natural language in ambient intelligence seems to be a reasonable bet. This makes it necessary for computer systems to understand natural language at the human proficiency level. If guessing games are to be performed, they have to be built on the platform of solid content and context comprehension to be useful for any reasonable purpose.

We assume, therefore, that an effective computer system that is used to communicate naturally with humans must be equipped with the same ability to learn new terms in a familiar context as the ability that humans possess and utilize under the same circumstances. It is also reasonable to assume that such learning is based on the perception of the situation described in the text, and that, as with most perception-based mechanisms, it is fuzzy.

II. WHAT KIND OF ANIMAL IS UNATTESTED INPUT?

A. First Type: Annetested Input

Annetested input, along with ambiguity, is a major problem for systems attempting to understand natural language: no

matter how many words the systems is supposed to be familiar with, there are hundreds of thousands words and senses that it has never encountered and is, therefore, not equipped to handle with the resources it has. However, in order to be robust, such a system must have such a capability.

Annetested input processing can only be meaningful when it happens among the words and in sentences that are well understood other than this one particular unknown occurrence. The task then becomes to guess the meaning of the annetested input on the basis of the fully understood information in the rest of the sentence. In a way, annetested input is being solved just like an inequality system in math is solved, by finding the possible intervals which satisfy it.

Non-meaning based approaches cannot proceed as far because they are not aware of the meaning of the other words or utterances in the sentence. The only reason the famous Jabberwocky poem [1] is understood somewhat by humans is that, besides the felicitous morphological and syntactic guesses, they fully understand quite a few words interspersed with the unknown ones.

The first stanza is, perhaps, the most artful one, because the only known words there are of a grammatical nature: conjunctions, prepositions, determiners and auxiliary verbs. As a result, the understanding is the vaguest and most surprising—in the sense of having any understanding at all:

Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

By contrast, in the most easily understood stanzas (third and fifth, respectively), the annetested input is limited to very few unknown non-proper-name words each:

He took his vorpal sword in hand:
Long time the manxome foe he sought –
So rested he by the Tumtum tree,
And stood awhile in thought.

One, two! One, two! And through and through
The vorpal blade went snicker-snack
He left it dead, and with its head
He went galumphing back.

Notice that the first occurrence of *vorpai* is broad: it can be any attribute of a sword. The second occurrence modifies the blade of a sword, thus reducing the set of possible senses of the annotated word. In a similar fashion, the fully familiar and well understood words in this section have compensated, we hope, for our deliberate misspelling of the word *unattested* as *anetested*.

B. The Vagueness of *Vorpai*

One leaves the Jabberwocky poem with a clear understanding that *vorpai* is an attribute of *blade*. This understanding is, however, vague in the sense of the coarsest possible grain size for a blade attribute: it can be such a physical property as sharp (primary function quality), steel (material), masterful (manufacturing execution), heavy (weight), long (dimension); it can also be a non-physical property, such as expensive (price), famous or legendary (extent of what is known about it), heroic (history of performance), merciless (effectiveness), etc. But we cannot yet treat the property as fuzzy because there is no basis for assigning a non-100% degree of membership to any one of the possibilities (cf. [2] on possibilistic theory and the Hartley measure of the uncertainty). So the captured meaning of *vorpai* is uncertain in the sense of vagueness but, perhaps, not yet fuzzy.

III. THE MEANING OF UNKNOWN WORDS

While the meaning of the sentence is not fully compositional in general, many sentence meanings can be put together from individual word meanings, provided that the syntax and semantics make sense. All words in the sentence are connected through some semantic relationships within the syntactic phrases to which they belong. We will assume that some two words in a syntactic phrase are connected through one or more relationships, and each word is connected to at least one other word in the sentence through such relationship. Moreover, there is a finite number of existing relationships. For simplicity, we will describe here binary relations that correspond to such semantic relationships. Such semantic relationships in natural language are fuzzy: we unlock a door with a key, with a code, or with a paper clip¹. Clearly, $\mu_{\text{INSTRUMENT}}(\text{UNLOCK}, \text{KEY}) < \mu_{\text{INSTRUMENT}}(\text{UNLOCK}, \text{PAPER-CLIP})$.

Each such relation $R(X, Y)$, where $x \in X$ and $y \in Y$, has a domain and range, defined respectively as:

$$\mu_{\text{DOM } R}(x) = \max_{y \in Y} \mu_R(x, y), \text{ for each } x \in X.$$

$$\mu_{\text{RAN } R}(y) = \max_{x \in X} \mu_R(x, y), \text{ for each } y \in Y.$$

Furthermore, we assume that each relation has at least some (x, y) such that $\mu_R(x, y) = 1$: $\max_{y \in Y} \max_{x \in X} \mu_R(x, y) = 1$.

Semantic relationships typically have well defined and easily calculated inverse relationships, with

$$\mu_{\text{DOM } R}^{-1}(y) = \max_{x \in X} \mu_{R^{-1}}(y, x), \text{ for each } y \in Y.$$

$$\mu_{\text{RAN } R}^{-1}(y) = \max_{x \in X} \mu_{R^{-1}}(y, x), \text{ for each } x \in X.$$

The relationships are typically verb centered: $\text{agent}(y, x_1)$, $\text{theme}(y, x_2)$, $\text{beneficiary}(y, x_3)$, etc. where y is a verb. Yet, they are rarely all interdependent and, hence, can be dealt with in pairs. When a relationship is object centered, the task is simpler, as the number of such relationships is smaller. As an example, consider the following sentence: *The man opened the door to his house with the key*. The following relationships are of interest: $\text{AGENT}(\text{OPEN}, \text{MAN})$, $\text{THEME}(\text{OPEN}, \text{DOOR})$, $\text{INSTRUMENT}(\text{OPEN}, \text{KEY})$, $\text{ENTRY-POINT}(\text{DOOR}, \text{HOUSE})$, $\text{OWNER}(\text{MAN}, \text{HOUSE})$. There are 3 verb centered relationships here, and 2 object centered ones, as shown in Fig. 1.

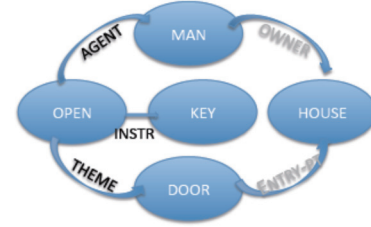


Figure 1. A linked representation of *The man opened the door to his house with the key*.

Whenever a sentence consists of compositional relationships between words, where the words are not interdependent, an unknown word can be calculated based on the domain or range of the relation that corresponds to the semantic relationship between the unknown and the known words. Thus, if in the example above, we don't know the word *man* but somehow know that it is an AGENT of OPEN, we can derive some knowledge of what *man* is from $\mu_{\text{AGENT}}(\text{OPEN}, \text{man})$. Moreover, until more information is found we can approximate *man* as something that maximizes $\mu_{\text{AGENT}}(\text{OPEN}, z)$.

On the other hand, there are situations when an unknown word depends on more than one relationship. This can happen in several situations, some of which are shown in Figure 1. For instance, if the word *house* were unknown, its meaning would be derived from something that can be owned and from something that has a door as an entry. Similarly, if the word *door* was unknown, its meaning would have to be derived from something that is a theme of OPEN and that serves as an entry to HOUSE.

Assuming that there are at least two relations $R1(X, Y)$ and $R2(Y, X)$ such that there is a $y \in Y$ where $(x, y) \in R1$ and $(y, z) \in R2$, a join can be formed between these two relations ($R1 * R2$), such as

$$\mu_{R1 * R2}(x, y, z) = \min[\mu_{R1}(x, y), \mu_{R2}(y, z)]$$

for each $x \in X, y \in Y, z \in Z$ ([2], p. 77). Using this formula, the meaning of unknown words—as those in the examples of *door* and *house*—can be calculated.

The formula can be generalized to n relations:

$$\begin{aligned} \mu_{R1 * R2 * \dots * Rn}(x_1, x_2, \dots, x_n, y) \\ = \min[\mu_{R1}(x_1, y), \mu_{R2}(x_2, y), \dots, \mu_{Rn}(x_n, y)] \end{aligned}$$

¹ http://www.ehow.com/how_4809693_unlock-locked-door-key.html.

001. ☒ e heap of scattered booty and fired. He dropped a MAN WITH THE first bullet. At the same moment Whee
 002. ☒ of pain. Brassnose was strangely silent. The big MAN WITH THE whitened hair murmured something: his
 003. ☒ d called to her liveried driver. This was the big MAN WITH THE proprietary air and the beetling, sha
 004. ☒ re afraid. Not so soon, anyway. I clapped the big MAN WITH THE bleached hair on his shoulder and sai
 005. ☒ h one of his windows- as though this broken-nosed MAN WITH THE muscular arms and wrestler's neck was
 006. ☒ ce that can happen to best of orchestras when the MAN WITH THE baton lacks technique and style. #BAY
 007. ☒ etimes fights his own evident charm), and a third MAN WITH THE engaging name of Birdwhistell (Ray),
 008. ☒ ors- whose pitcher was a formidable looking young MAN WITH THE only baseball cap. This was a bitterl

Figure 2. Results for *man with the* ____

001. ☐ try. His bandages seemed on fire. He had shut his DOOR WITH THE brass number screwed to it. In the k

Figure 3. Results for *door with the* ____

for each $x_i \in X_i$, $y \in Y$. Using this formula, if word *open* is unknown, its meaning can be derived by considering together its AGENT *man*, its INSTRUMENT *key* and its THEME *door*.

The situation gets more complicated when we consider unknowns that are dependent on other information from the sentence. For example, if we have *A cat opened the door with xyz*, we know that typical instruments of opening the door like keys and code are excluded. On the other hand, if it is a locksmith that opened the door, it is probably not the regular key that was used to open it. Intuitively, in such cases the join should not be done on a single concept but on as many as it takes to derive the correct meaning. Thus, in the case of *cat* or *locksmith*, the agent has to be considered on the join as well or membership functions has to be defined based on individual agents. We will abandon the notion of individual membership function of the INSTRUMENT of OPEN based on the agency for a time being, but note that it is not much different than defining sets for the height of a basketball player: it is easy to define them, but they are not practical from the point of view of natural language usage.

It is easy to see that for the cases of unknown words that are independent arguments of a binary relation, their meaning can be derived from a join of all pertinent relations contained throughout a text: *I lost the xyz that opened the door last time. It fell through the opening in front of the elevator door.* Here, not only do we know that *xyz* is something that opened the door, but it is also slim enough to fall through a centimeter (or so) wide opening. Thus, if somebody disambiguates a sentence, and assigns arguments to each relation, including which relation an unknown word will be in, it is mathematically possible to find the meaning of such word, especially, if it is used in many sentences that describe different functional characteristics. The next section will demonstrate what type of linguistic processes are needed do it automatically.

IV. WHAT DOES IT TAKE TO OPEN A DOOR?

Let us, again, consider the sentence *The man opened the door with the xyz*. In the sentence, *xyz* can be a number of things, including *key*, *combination code*, *card*, *two hanger wires*, or any other instrument with the use of which a lock on the door can be opened. Perhaps somewhat less obviously, *xyz* can be *mirror*, *picture*, *hole*, *ornament*, or any other detail or part of the door. Even less obviously, and somewhat syntactically awkward, *xyz* may stand for *friend*, *dog*, or any other accompanier for *the man*.

Without the explanation above, the sentence is three-way syntactically ambiguous, depending on the analysis of the prepositional phrase which may pertain to *man*, *open*, or *door*. Thus, similarly to the case of *vorpai*, indicated are the three vague sets of constituents that somehow modify the dominant entities above.

The task of unattested input processing is, then, to determine the semantic features that can satisfy *the man with the xyz*, *open with the xyz*, or *the door with the xyz*. It is here that understanding these phrases and the resulting sentence is necessary. The process of understanding the phrases when all words are known is outlined in Section V.

It is tempting to use the corpus-based approach and look for occurrences of *xyz* given *the man with the*, *open with the*, *the door with the xyz*. A brief search² of the Brown corpus for the first phrase yielded the results shown in Fig. 2. The results of the second phrase produced nothing, the results of the third phrase yielded one usage, shown in Fig. 3.

The lessons of this example are fast learned: none of the usages received for the first query work for the sentence *The man opened the door with the xyz*. The one usage in the third query is extremely limiting: there are many more possibilities for the detail or part of the door other than having a brass number.

It can be argued that the exact phrase search approach is too narrow and should be expanded to the conceptual search³, at least of the dominant word. It is possible that it will somewhat enrich the results and give a somewhat higher quantity of the acceptable phrases, such as and beyond *brass number*. It would not, however, correct the situation with the first tried phrase, illustrated in Fig. 2, where none of the results were suitable for the example sentence.

Nirenburg & Raskin [3] sketch an early meaning-based procedure that can be used to find the prevalent crisp set associated with the sense of the one unknown word in a sentence. (See also references there and elsewhere on pre-semantic attempts to account for unattested input: for a representative sample, see [4-19]) Using a much simplified version of early Ontological Semantics, they demonstrate that non-existing word *keegh* can be understood as an instrument of locking/unlocking a door on the basis of the machine-

² <http://conc.lectutor.ca/concordancers> was used

³ a primitive example of which could be the usage of hypernyms and hyponyms of WordNet.

understood definitions of *lock*, *door*, and the ontological knowledge of the most common artifact that is required to perform the unlocking.

Such an approach, especially on a single sentence, demonstrates that a solution can be found, without validating that it was the optimal solution (see [20] for empirical verification). Consider, for example, a sentence *he unlocked the car door with an xyz*. The range of possibilities here encompasses *key*, *keyless remote fob*, *code* (on the door), *wire*, *screwdriver*, *hanger*, *lock pick* (*Slim Jim*). In a crisp treatment, *key*, *keyless remote fob*, *code*, or maybe *lock pick* and *hanger* will work. The *screwdriver*, on the other hand, will probably draw a blank.

Now suppose, the sentence that follows is *He always had the xyz with him to take the screws off laptops*. The screwdriver, then, an initially unlikely tool, becomes the leading candidate for the unattested input: while it is possible, although hard, to unlock a car door with a screwdriver, none of the other tools are purposefully used to take the screws off laptops.

It appears useful, then, to collect all interpretations, to the lowest possible degree (as unlikely as they may sound), of the tools that could imaginably open a car door. The set of all these interpretations will remain intact until and unless further information is perceived to narrow it down. If decisions have to be made prior to the moment when new, narrowing information is received, then it is natural to accept those tools that have the highest membership in the set as the interpretation of *xyz*. Moreover, it seems natural to select the most probable of them to be used for the interpretation as well. Thus, [3] concluded—or rather, stopped just short of concluding—that the unattested input was an instrument for opening a door, the likeliest of which is, of course, some sort of key.

V. A MEANING-BASED METHOD FOR GUESSING WHAT XYZ IS

In this section, we will demonstrate in somewhat simplified non-proprietary technical detail, how the meaning of an unknown word is determined on the basis of the full understanding of everything else in the sentence. The derived meaning will be calculated based on the fuzzy joins of relevant relations. This will be done with the methods and resources of the Ontological Semantic Technology (OST—see [21-23]).

At the core of OST are repositories of world and linguistic knowledge, acquired semi-automatically within the approach and used to disambiguate the different meanings of words and sentences and to represent them. These repositories, also known as the static knowledge resources, consist of the ontology, containing language-independent concepts and relationships between them; one lexicon per supported language (English, Russian, etc), containing word senses anchored in the language-independent ontology which is used to represent their meaning; the Proper Name Dictionary (PND), which contains names of people, countries, organizations, etc., and their description anchoring them in ontological concepts and interlinking them with other PND entries; and a common sense rules resource. A conforming lexicon and ontology, as well as – when needed -- PNDs and common sense rules, are

used by the Semantic Text Analyzer (STAn), a software, developed by RiverGlass Inc., that produces Text Meaning Representations (TMRs) from the text that it reads. The format of TMRs conforms to the format and interpretation of the ontology. The processed TMRs are entered into InfoStore, a dynamic knowledge resource of OST, from which information is used for further processing and reasoning.

A. Lexicon

The substantive part of interpreting the sentence *The man opened the door with the xyz* starts with the search for the event of the single clause, and in this case, it is easy: *open* is the only word with any sense anchored in an ontological event concept. For illustrative purposes, we will work with the following senses of *open* as a verb that a semantic analyzer would find in a generic OST lexicon:

1. to allow access [*open the door*]
2. spread out, unfold (open out) [*open the wings*]
3. allow public access, permit access [*open the gallery to the public*]
4. establish [*open a new restaurant*]
- ...

The formal definition of the above senses is:

```
(open-v1
  (cat(v))
  ...
  (syn-struct((subject((root($var1))(cat(np))))
    (root($var0))(cat(v))
    (directobject((root($var2))(cat(np))))
    (pp-adjunct((opt(+))(root(with by))(cat(prepp))
      (obj((root($var3))(cat(np))))))
    (pp-adjunct((opt(+))(root(to into of))(cat(prepp))
      (obj((root($var4))(cat(np))))))
  ))
  (sem-struct(open
    (agent(value(^$var1)))
    (theme(value(^$var2)))
    (instrument(value(^$var3)))
    (location(value(^$var4)))
  ))
)
(open-v2/v3/v4
  (cat(v))
  ...
  (syn-struct((subject((root($var1))(cat(np))))
    (root($var0))(cat(v))
    (phr((opt(+))(root(out up))(cat(prepp))))
    (directobject((root($var2))(cat(np))))
    (pp-adjunct((opt(+))(root(for to))(cat(prepp))
      (obj((root($var3))(cat(np))))))
  ))
  (sem-struct(unfold/give-access/establish
    (agent(value(^$var1)))
    (theme(value(^$var2)))
    (beneficiary(value(^$var3)))
  ))
)
```

For the simplicity purposes, we will consider only one (physical) sense of *door* (physical movable door), only one sense of *man*, and the following, non phrasal, senses of *with*:

1. accompanier (animate or thing) [*went with a friend*]

2. instrument [*cut with a knife*]
3. beneficiary [*left the parcel with Bob*]
4. object-as-part [*door with a handle*]
5. location-of [*door with a mirror*]
6. containment [*bowl with cherries*]
7. manner [*ate with gusto*]

The first six definitions of *with* can be represented in the following format, where PROP is the property from the list above:

```
(with-prep[1-3]
...
(syn-struct((vp((root($var1))(cat(vp))))
  (root($var0))(cat(preop))
  (obj((root($var2))(cat(np))))))
(sem-struct(^$var1(PROP(value(^$var2))))))
)
(with-prep[4-6]
...
(syn-struct((np((root($var1))(cat(np))))
  (root($var0))(cat(preop))
  (obj((root($var2))(cat(np))))))
(sem-struct(^$var1(PROP(value(^$var2))))))
)
```

The last sense is defined as:

```
(with-prep7
...
(syn-struct((vp((root($var1))(cat(vp))))
  (root($var0))(cat(preop))
  (obj((root($var2))(cat(np adv))))))
(sem-struct(^$var1(manner(value(^$var2))))))
)
```

The same variables in the syntactic and semantic structure (syn-struct and sem-struct) are bound together, with ‘^’ standing for the meaning of the variable that follows the symbol, and ‘\$’ standing for the value of the variable. In other words, ‘\$’ refers to a word that a variable points to, and ‘^’ to its ontological meaning. Because the lexicon senses only point to the conceptual definitions in the ontology, the bulk of work is carried by the ontology.

B. Ontology

The ontology is a lattice that contains the knowledge necessary to accept or reject an interpretation of a sentence. Each concept is defined through properties and their fillers. Each filler is accented by the facet which is interpreted as supplying the degree of membership in a set.

An ontological concept has the following format:

```
concept-name
  (attribute(facets(literal-value+))+)*
  (relation(facet(concept-name+))+)*
facet4
  sem | value | default | relaxable-to | not
```

In this paper, we are only interested in relations. The following interpretations for ontological definitions containing fuzzy concepts and relations were defined in [24]:

Given a set of objects $\mathcal{D}$ and given its interpretation function I , for every fuzzy concept B , object x is an element of B with some degree $I[B](x) \rightarrow [0, 1]$; for every relation Rel , $I[Rel](x, y) \subseteq \mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$; We will assume that $x \in B$ if $I[B](x) \rightarrow (0, 1]$.

With this in mind, the following is true for C and D , using [15]:

$$\begin{aligned}
 I[C \ D](x) &= \max\{I[C](x), I[D](x)\} \\
 I[\text{and } C \ D](x) &= \min\{I[C](x), I[D](x)\} \\
 I[(\text{Rel}(D))](x) &= \max_{y \in D}\{I[Rel](y, x)\} \\
 I[(\text{Rel}(\text{and } C \ D))](x) &= \min\{I[Rel(C)](x), I[Rel(D)](x)\} \\
 I[Rel(C \ D)](x) &= \max\{I[Rel(C)](x), I[Rel(D)](x)\} \\
 I[C(Rel(D))](x) &= \min\{I[C](x), I[Rel(D)](x)\} \\
 I[C(Rel_1(D))(Rel_2(E))](x) \\
 &= \min\{I[C(Rel_1(D))](x), I[C(Rel_2(E))](x)\}
 \end{aligned}$$

For the construction of the ontology, subsumption definition of [25] was followed: C subsumes D iff $\forall x \in \mathcal{D}$: $I[C](x) \geq I[D](x)$.

Fuzzy values of facets as the degree of membership is assumed to be set by the ontology engineer, with DEFAULT facet always corresponding to 1, NOT facet corresponding to 0, while the other three facets varying depending on the application and the domain. We will use a slightly different notation in the paper, specifying the exact degree of membership instead of using a facet. Thus, we adopt the following format, where membership-degree is any number between 0 and 1:

```
concept-name
  (relation(membership-degree(concept-name+))+)*
```

We assume that a parent concept has at most the same membership degree value as a child concept (unless inheritance is blocked). Then, given a concept B that is a descendant of a concept A and an ancestor of a concept C , where $\mu_P(A)$ and $\mu_P(C)$ are known:

$$\mu_P(B) = \mu_P(A) + \frac{\mu_P(C) * \text{dist}(A, B)}{\text{dist}(A, C)},$$

where $\text{dist}(X, Y)$ is the logarithmic distance between the lattice nodes X and Y and either $\text{ancestor}(X, Y)$ or $\text{ancestor}(Y, X)$ holds.

The somewhat simplified ontological concepts OPEN, UNFOLD, GIVE-ACCESS, and ESTABLISH for the 4 lexical senses of *open* above and the concept DOOR for *door* are shown below, without their inherited properties, unless access to them is warranted by the analyzed sentence:

```
(open
  (definition(...("to open a container door window etc")))
  (is-a(1(change-position)))
  (theme(1(container door window))(0.75(artifact))
    (0.1(physical-object)))
  (instrument(1(key))(0.75(artifact)))
  (agent(1(human))(0.5(animal)))
  (accompanier(1(human))(0.75(animal))))
)
(unfold
```

⁴This list of facets is only applicable to concepts

```

    (definition(...("to change configuration of parts of an object")))
    (is-a(1(change-configuration)))
    (theme(0.75(artifact(part-of-object(0.5(physical-object))))
    (instrument(0.9(body-part))(0.5(artifact))))
  )
  (give-access
    (definition(...("to allow public access")))
    (is-a(1(permit)))
    (theme(1(building))(0.5(organization))
      (0.1(geopolitical-entity)))
    (beneficiary(0.75(human))(0.5(organization))(0.1(animal)))
    (precondition(0.75(communicative-act)))
  )
  (establish
    (definition(...("to establish an institution")))
    (is-a(1(create-social-object)))
    (theme(0.5(group))(0.65(organization))
      (0.1(geopolitical-entity)))
  )
  (movable-door
    (definition(...("the slab that is open and closed as opposed to
the hole in the wall")))
    (is-a(1(building-fixture)))
    (inside-of(1(door)))
    (contains(0.05(physical-object)))
    (has-object-as-part(1(handle))(0.8(lock)))
    (locale-of(0.5(mirror hook writing-symbol))(0.3(ornament)))
  )
  (man
    (definition(...("the slab that is open and closed as opposed to
the hole in the wall")))
    (is-a(1(human)))
    (gender(1(male)))
    (age(1(adult))(0.75(teenager)))
    (has-object-as-part(1(male-body-parts)))
    (locale-of(0.75(garment accessory)))
  )

```

In a situation when a word is unknown, humans use the linguistic clues from the sentence as well as their knowledge of the world to come up with some approximate interpretation of the unknown word. The linguistic clues consist of tentative morphological, syntactic and semantic information (such as the lexical senses of the known words), while the knowledge of the world for humans can range from purely encyclopedic to common sense. The ontology tries to capture parts of such knowledge, also relying on InfoStore for facts and common sense rules resources for the rest.

Similarly, a computer calculates the same possibilities, with the membership functions reflecting the semantic preferences. Thus, when the phrase *open the door* is processed, all but the open-v1 sense are excluded due to the membership value of 0 of some phrase interpretations:

```

 $\mu_{AGENT}(OPEN, MAN)=1$ ;  $\mu_{THEME}(OPEN, DOOR)=1$ ;
 $\mu_{THEME}(UNFOLD, DOOR)=0$ ;  $\mu_{THEME}(GIVE-ACCESS, DOOR)=0$ ;
 $\mu_{THEME}(ESTABLISH, DOOR)=0$ ;

```

In other words, according to the ontology, it is not possible to unfold a door, allow public access to a door, or to establish a door as an institution.

C. Back to door with, man with or open with... xyz

Our sentence is *The man opened the door with the xyz*. The last word before xyz is the article, since there is no other word

after it, we can conclude that it is a noun. Thus information is relevant when we create the lexical sense of a word, but not in guessing its meaning in a sentence. We thus continue to the next closest word, which is the preposition *with*. Now we have more clues to work with. According to the lexical senses of *with*, xyz can be one of the following:

1. It can accompany an animate being in any event
2. It can be used as an instrument for a physical event that has a theme.
3. It can be a receiver of something from somebody in an exchange possession or communicative event.
4. It can be a part of a physical object
5. It can be physically located on an (physical) object
6. It can be contained within a physical object
7. It can signify a manner of an event

According to the syntactic structures, we know that the above definition have to conform to the following format in its normalized form:

1. The subject, indirect object and the predicate must be present in the sentence. The subject must precede the verb, while the prepositional phrase containing the indirect object can be in any position.
2. The prepositional phrase containing the instrument must be in the initial or final position, thus preceding or following the predicate with the direct object.
3. The prepositional phrase containing the beneficiary (receiver) must follow the direct object, which in turn follows the predicate.
4. The prepositional phrase must follow a noun (phrase).
5. The prepositional phrase must follow a noun (phrase).
6. The prepositional phrase must follow a noun (phrase).
7. The prepositional phrase must be used adverbially, contain no article, and follow the predicate, not necessarily immediately.

Applying these rules to the sentence, we can rule out *with-prep7* on a purely syntactic ground and *with-prep3* since *open* has no sense that corresponds to the exchange possession event or communicative event on a purely semantic ground.

An interesting case is presented by *with-prep6*: while it is possible in principle to fill the inside of a door (or even an a man's body part) with something, their $\mu_{CONTAINS}$ is very low. Thus, while they are possible and the sense is not discarded, they are unlikely to become serious candidates contributing to the sense of xyz unless some other information props them up. In the case of *door* as a container filled with xyz,

$\mu_{CONTAINS}(MOVABLE-DOOR, xyz)=0.05$,

where xyz is any PHYSICAL-OBJECT small enough to fit inside a hollow part of a door, if any.

The cases of *with-prep4,5* remain valid. The difference between *with-prep4* and *with-prep5* is that the former realized only when the following noun is a part of something while the latter is realized when the following noun is located on something. The following interpretations are then valid:

$\mu_{LOCALE-OF}(MOVABLE-DOOR, xyz)=0.5$,

where xyz is any descendant of MIRROR, HOOK, WRITING-SYMBOL;

$\mu_{\text{LOCALE-OF}}(\text{MOVABLE-DOOR}, xyz)=0.3$,
 where xyz is any descendant of ORNAMENT;
 $\mu_{\text{HAS-OBJECT-AS-PART}}(\text{MOVABLE-DOOR}, xyz)=1$,
 where xyz is any descendant of HANDLE;
 $\mu_{\text{HAS-OBJECT-AS-PART}}(\text{MOVABLE-DOOR}, xyz)=0.8$,
 where xyz is any descendant of LOCK.

In theory,

$\mu_{\text{LOCALE-OF}}(\text{MAN}, xyz)=0.75$,
 where xyz is any descendant of GARMENT or ACCESSORY;
 $\mu_{\text{HAS-OBJECT-AS-PART}}(\text{MAN}, xyz)=1$,

where xyz is any descendant of MALE-BODY-PARTS. However, the strongly prescriptive direct word order in English makes it extremely syntactically unlikely that the prepositional phrase at the end of the sentence would modify *man* ($\mu_{\text{SYNTAX}}(\text{man}, xyz)=0.05$), thus, lowering this possibility to practically non-existent. Unlike the case of *with-prep6*, where additional semantic information could prop the sense up, it is considerably less likely for such propping syntactic information to be found.

Also valid are the cases of *with-prep1,2*. The only event that can be considered with these two senses of *with* is OPEN, as UNFOLD, GIVE-ACCESS and ESTABLISH were excluded above due to their incompatibility with *door* (the corresponding concept, MOVEABLE-DOOR).

We will consider *with-prep1* first. According to its syntactic and semantic structures, it needs a subject that is an agent of an event that can be accompanied by an animate object that is part of the prepositional phrase. Thus, *man* is an agent of *open* and xyz must then be an accompanier of *man* in this event:

$\mu_{\text{ACCOMPANIER}}(\text{MAN}, xyz)=1$,
 where xyz is HUMAN, and
 $\mu_{\text{ACCOMPANIER}}(\text{MAN}, xyz)=0.75$,
 where xyz is ANIMAL.

In the case of *with-prep2*, xyz must be an instrument of opening the door. Thus,

$\mu_{\text{INSTRUMENT}}(\text{OPEN}, xyz)=1$,
 where xyz is KEY, and
 $\mu_{\text{INSTRUMENT}}(\text{OPEN}, xyz)=0.75$,
 where xyz is any other ARTIFACT.

Similarly to the case of *with-prep6*, we now have to consider the word order in the resulting interpretations in relation to the prescriptive and preferred English sentence word order. Thus, we have to use yet another membership function, that corresponding to the perceived syntactic appropriateness or smoothness.

$\mu_{\text{SYNTAX}}(\text{The man opened the door with the dog}) <<$
 $\mu_{\text{SYNTAX}}(\text{The man with the dog opened the door})$

In more general terms,

$\mu_{\text{SYNTAX}}(\text{AGENT EVENT THEME with ACCOMPANIER})=0.2 <<$
 $\mu_{\text{SYNTAX}}(\text{AGENT with ACCOMPANIER EVENT THEME})=1$

Thus, we have the following possible interpretations of xyz , all with different degrees of acceptability of the alternative sentences: mirror, hook, writing-symbol; ornament; handle; lock; garment, accessory; male-body-parts; human; animal; key; artifact. The acceptability depends both on the semantics and syntax of the resulting sentences.

VI. PUTTING IT ALL TOGETHER: CALCULATING DEGREES OF ACCEPTABILITY

We define an interpretation as acceptable if and only if it is acceptable both syntactically and semantically. We treat such acceptability as the intersection of the syntactic and semantic sets.

Syntactic acceptability of a sentence is defined as a minimum acceptability of all of its (binarized) phrases; where the acceptability of each phrase is maximized to the most acceptable syntactic analysis in case of syntactic ambiguity:

$$\mu_{\text{SYNTAX}}(\text{sent}) = \min_{\text{phr} \in \text{sent}} \max_{x,y \in \text{phr}} [\mu_{\text{phr}}(x,y)]$$

Semantic acceptability is similarly defined as a minimum of acceptability of all semantic relations on the maximally acceptable interpretation of each relation:

$$\mu_{\text{SEMANTICS}}(\text{sent}) = \min_{R \in \text{sent}} \max_{x,y \in R} [\mu_R(x,y)]$$

The overall syntactic/semantic acceptability of a sentence is, then:

$$\mu_{\text{ACCEPTABILITY}}(\text{sentence}) = \min[\mu_{\text{SYNTAX}}, \mu_{\text{SEMANTICS}}]$$

In *The man opened the door with the xyz*, we have eleven possible components that vary with different meanings of xyz :

$$\mu_{\text{ACCEPTABILITY}}([\text{very small}] \text{ PHYS-OBJECT } [\text{contained in door}]) = \min\{1, 0.05\} = 0.05$$

$$\mu_{\text{ACCEPTABILITY}}(\text{MIRROR, HOOK, WRITING-SYMBOL } [\text{on door}]) = \min\{1, 0.5\} = 0.5$$

$$\mu_{\text{ACCEPTABILITY}}(\text{ORNAMENT } [\text{on door}]) = \min\{1, 0.3\} = 0.3$$

$$\mu_{\text{ACCEPTABILITY}}(\text{HANDLE } [\text{part of door}]) = \min\{1, 1\} = 1$$

$$\mu_{\text{ACCEPTABILITY}}(\text{LOCK } [\text{part of door}]) = \min\{1, 0.8\} = 0.8$$

$$\mu_{\text{ACCEPTABILITY}}(\text{GARMENT, ACCESSORY } [\text{of man}]) = \min\{0.05, 0.75\} = 0.05$$

$$\mu_{\text{ACCEPTABILITY}}(\text{BODY-PART } [\text{of man}]) = \min\{0.05, 1\} = 0.05$$

$$\mu_{\text{ACCEPTABILITY}}(\text{HUMAN } [\text{accompanies man}]) = \min\{0.2, 1\} = 0.2$$

$$\mu_{\text{ACCEPTABILITY}}(\text{ANIMAL } [\text{with man}]) = \min\{0.2, 0.75\} = 0.2$$

$$\mu_{\text{ACCEPTABILITY}}(\text{KEY } [\text{as instrument}]) = \min\{1, 1\} = 1$$

$$\mu_{\text{ACCEPTABILITY}}(\text{ARTIFACT } [\text{instrument}]) = \min\{1, 0.75\} = 0.75$$

Thus, the two interpretations of the sentence, where xyz results in the highest syntax-semantic acceptability, are:

- *The man opened the door with the handle.*
- *The man opened the door with the key.*

While it appears to be intuitively correct, in neither interpretation is it really necessary to communicate the prepositional phrase. We propose an interesting observation: things that have the membership value of 1 are there by default, as it were, and are likely to be dropped from the sentence. The exception is when these defaults are modified: it is unlikely that

we would hear *he unlocked the door with a key* – what else can he unlock the door with by default? – but it is ok to say *he unlocked the door with a golden key* or *spare key*, or *his friend's key*.

Thus, if we wish to model human communication, we should be more likely to consider concept candidates for the unknown words that are a little less than 1 than those at exactly 1. On the other hand, the grammaticality of a sentence should not be compromised. Thus, if a verb requires a direct object, and the unknown word is this direct object, then preferring the membership value of less than 1 should not play into the decision process—as the unknown word must be in the sentence for it to be grammatically correct. It is plausible that such transitive verbs develop an intransitive variant, in which case the preferred membership of exactly 1 consideration is no longer valid: it is then so preferred that it becomes trivial and must be dropped.

More generally, acceptability, which is typically barely mentioned in linguistics, cannot be explored scientifically without a serious fuzzy logic influx, and the math, tentatively offered here, may have to be developed further to capture the individual dimensions of this phenomenon as well as combining them together. Besides the syntactic and semantic dimensions that we have touched upon here, there are the:

- appropriateness dimension, which renders the sentence, *This is the best night of my life*, with the membership degree of 1 on both the syntactic and semantic dimensions, totally inappropriate (or humorous, metaphoric, or poetic) when uttered in the middle of the day;
- relevance dimension that demands that each sentence in a text be clearly related to the surrounding sentences or see its membership degree sharply decreased;
- saliency dimension, where, again, a sentence with the degree membership of 1 on the other dimensions may have a much lower value if it does not convey new and important information.

And the list is not exhaustive because, for instance, humorous, metaphoric, and poetic are probably also additional dimensions. While the membership functions will be more difficult to define in the current state of scholarship on those dimensions, the principle of calculating the acceptability of a sentence, which in turn helps with determining meanings of unknown words, will—we believe—remain essentially unchanged.

VII. REFERENCES

- [1] L. Carroll, *Through the Looking-Glass and What Alice Found There*. 1872.
- [2] G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [3] S. Nirenburg and V. Raskin, *Ontological Semantics*. Cambridge, MA: MIT Press, 2004.
- [4] J. Smarr and C. D. Manning, “Classifying unknown proper noun phrases without context,” in: Technical Report dbpubs/2002-46, NLP Group, Stanford University, 2002.
- [5] D. Bikel, R. Schwartz, and R. Weischedel, “An algorithm that learns what's in a name,” *Machine Learning* 1999, 34:pp. 211–231.
- [6] A. Mikheev, “Automatic rule induction for unknown word guessing,” *Computational Linguistics* 1997, 23(3), pp. 405–423.
- [7] M. Ciaramita and M. Johnson, “Supersense tagging of unknown nouns in WordNet,” in: *Proceedings of EMNLP-2003*, pp. 594–602.
- [8] J. Curran, “Supersense tagging of unknown nouns using semantic similarity,” in: *Proceedings of ACL-2006*, pp. 26–33.
- [9] Collins, M. and Y. Singer, “Unsupervised models for named entity classification,” in: *Proceedings of EMNLP/VLC'99*, P. Fung and J. Zhou (Eds.), College Park, MD: ACL, 1999.
- [10] S. Cucerzan and D. Yarowsky, “Language independent named entity recognition combining morphological and contextual evidence,” in: *Proceedings of EMNLP/VLC'99*, P. Fung and J. Zhou (Eds.), College Park, MD: ACL, 1999.
- [11] S. Buchholz and A. Van Den Bosch, “Integrating seed names and n-grams for a named entity list and classifier,” in: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.
- [12] J. R. Curran, “Supersense tagging of unknown nouns using semantic similarity,” in: *Proc. of ACL*, Ann Arbor, MI: University of Michigan, 2005, pp. 26–33.
- [13] Murawaki, Yugo and Sadao Kurohashi, “Online acquisition of Japanese unknown morphemes using morphological constraints,” in: *Proc. of EMNLP*, 2008, pp. 429–437.
- [14] R. M. Weischedel, M. Meteer, R. M. Schwartz, L. A. Ramshaw, J. Palmucci, “Coping with ambiguity and unknown words through probabilistic models,” *Computational Linguistics*, 1993, 19(2): pp. 359–382.
- [15] M. Ciaramita, “Boosting automatic lexical acquisition with morphological information,” in: *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, Philadelphia, PA, USA, 2002, pp. 17–25.
- [16] D. A. Campbell and S. B. Johnson, “A technique for semantic classification of unknown words using UMLS resources,” in: *Proc AMIA Symp.* 1999: pp. 716–720.
- [17] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” in: *Named entities: Recognition, classification and use*, S. Sekine & E. Ranchhod (Eds.), Amsterdam: John Benjamins, 2009, pp. 3–25.
- [18] Y. Wilks, “Making preferences more active,” *Artificial Intelligence* 11(3), 1972, pp. 197–223.
- [19] Y. Wilks, and R. Catizone, “Lexical tuning,” *Proceedings of Computational Linguistics and Intelligent Text Processing*, Third International Conference (CICLING 2002), Mexico City, Mexico. 2002, pp. 105–126.
- [20] J. M. Taylor, V. Raskin, and C. F. Hempelmann, “Towards computational guessing of unknown word meanings: The ontological semantic approach,” submitted to CogSci 2011.
- [21] V. Raskin, C. F. Hempelmann, and J. M. Taylor, “Guessing vs. knowing: The two approaches to semantics in natural language processing,” A. E. Kibrik (Ed.), *Proceedings of Annual International Conference Dialogue*, Moscow, Russia, AABBY/Yandex, 2010.
- [22] J. M. Taylor, V. Raskin, V., and C. F. Hempelmann, “On an automatic acquisition toolbox for ontologies and lexicons in Ontological Semantics,” *International Conference on Artificial Intelligence*, Las Vegas, NE. 2010..
- [23] C. F. Hempelmann, J. M. Taylor, and V. Raskin, “Application-guided ontological engineering,” *Proceedings of International Conference on Artificial Intelligence*, H. A. Arabnia, D. de la Fuente, E. B. Kozzerenko, and J. A. Olivas (Eds.), Las Vegas, NE, 2010.
- [24] J. M. Taylor and V. Raskin, “Fuzzy Ontology for Natural Language,” *Proceedings of NAFIPS 2010*, Toronto, Canada, 2010
- [25] J. Yen, “Generalizing term subsumption languages to fuzzy logic,” *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence (IJCAI-91)*, 1991, pp. 472–477.