A Lexicon for Underspecified Semantic Tagging

Paul Buitelaar
Dept. of Computer Science
Brandeis University
Waltham, MA 02254-9110, USA
paulb@cs.brandeis.edu

Abstract

The paper defends the notion that semantic tagging should be viewed as more than *disambiguation* between senses. Instead, semantic tagging should be a first step in the interpretation process by assigning each lexical item a representation of *all* of its systematically related senses, from which further semantic processing steps can derive discourse dependent interpretations. This leads to a new type of semantic lexicon (CoreLex) that supports *underspecified semantic tagging* through a design based on *systematic polysemous classes* and a class-based acquisition of lexical knowledge for specific domains.

1 Underspecified semantic tagging

Semantic tagging has mostly been considered as nothing more than *disambiguation* to be performed along the same lines as part-of-speech tagging: given n lexical items each with m senses apply linguistic heuristics and/or statistical measures to pick the most likely sense for each lexical item (see eg: (Yarowsky, 1992) (Stevenson and Wilks, 1997)).

I do not believe this to be the right approach because it blurs the distinction between 'related' (*systematic polysemy*) and 'unrelated' senses (*homonymy* : bank - bank). Although homonyms need to be tagged with a disambiguated sense, this is not necessarily so in the case of systematic polysemy. There are two reasons for this that I will discuss briefly here.

First, the problem of multiple reference. Consider this example from the BROWN corpus:

> [**A long book heavily weighted with military technicalities**]$_{NP}$, in this edition it is neither so long nor so technical as it was originally.

The discourse marker (it) refers back to an NP that expresses more than one interpretation at the same time. The head of the NP (book) has a number of systematically related senses that are being expressed simultaneously. The meaning of book in this sentence cannot be disambiguated between the number of interpretations that are implied: the informational content of the book (**military technicalities**), its physical appearance (**heavily weighted**) and the events that are involved in its construction and use (**long**).

The example illustrates the fact that disambiguation between *related* senses is not always possible, which leads to the further question if a discrete distinction between such senses is desirable at all. A number of researchers have answered this question negatively (see eg: (Pustejovsky, 1995) (Killgariff, 1992)). Consider these examples from BROWN:

(1)	**fast**	run-up (of the stock)
(2)	**fast**	action (by the city government)
(3)	**fast**	footwork (by Washington)
(4)	**fast**	weight gaining
(5)	**fast**	condition (of the track)
(6)	**fast**	response time
(7)	**fast**	people
(8)	**fast**	ball

Each use of the adjective 'fast' in these examples has a slightly different interpretation that could be captured in a number of senses, reflecting the different syntactic and semantic patterns. For instance:

1.	**'a fast action'**	(1, 2, 3, 4)
2.	**'a fast state of affairs'**	(5, 6)
3.	**'a fast object'**	(7, 8)

On the other hand all of the interpretations have something in common also, namely the idea of 'speed'. It seems therefore useful to *underspecify* the lexical meaning of 'fast' to a representation that captures this primary semantic aspect and gives a general structure for its combination with other lexical items, both locally (in compositional semantics) and globally (in discourse structure).

Both the *multiple reference* and the *sense enumeration* problem show that lexical items mostly have an indefinite number of related but highly discourse dependent interpretations, between which cannot be distinguished by semantic tagging alone. Instead, semantic tagging should be a first step in the interpretation process by assigning each lexical item a representation of all of its systematically related 'senses'. Further semantic processing steps derive discourse dependent interpretations from this representation. Semantic tags are therefore more like *pointers* to complex knowledge representations, which can be seen as *underspecified* lexical meanings.

2 CoreLex: A Semantic Lexicon with Systematic Polysemous Classes

In this section I describe the structure and content of a lexicon (CoreLex) that builds on the assumptions about lexical semantics and discourse outlined above. More specifically, it is to be 'structured in such a way that it reflects the lexical semantics of a language in systematic and predictable ways' (Pustejovsky, Boguraev, and Johnston, 1995). This assumption is fundamentally different from the design philosophies behind existing lexical semantic resources like WordNet that do not account for any regularities between senses. For instance, WordNet assigns to the noun book the following senses:

publication
product, production
fact
dramatic_composition, dramatic_work
record
section, subdivision
journal

Figure 1: WordNet senses for the noun book

At the top of the WordNet hierarchy these seven senses can be reduced to two unrelated 'basic senses': the content that is being communicated (**communication**) and the medium of communication (**artifact**). More accurately, book should be assigned a qualia structure which implies both of these interpretations and connects them to each of the more specific senses that WordNet assigns: that is, **facts, drama** and a **journal** can be *part-of* the content of a book; a **section** is *part-of* both the content and the medium; **publication, production** and **recording** are all *events* in which both the content and the medium aspects of a book can be involved.

An important advantage of the CoreLex approach is more consistency among the assignments of lexical semantic structure. Consider the senses that WordNet assigns to door, gate and window:

door

movable_barrier	⇝ artifact
entrance	⇝ opening
access	⇝ cognition, knowledge
house	⇝ ??
room	⇝ ??

gate

movable_barrier	⇝ artifact
computer_circuit	⇝ opening
gross_income	⇝ opening

window

opening	⇝ opening
panel	⇝ artifact
display	⇝ cognition, knowledge

Figure 2: WordNet senses for the nouns door, window and gate

Obviously these are similar words, something which is not expressed in the WordNet sense assignments. In the CoreLex approach, these nouns are given the same semantic type, which is *underspecified* for any specific 'sense' but assigns them consistently with the same *basic* lexical semantic structure that expresses the regularities between all of their interpretations.

However, despite its shortcomings WordNet is a vast resource of lexical semantic knowledge that can be mined, restructured and extended, which makes it a good starting point for the construction of CoreLex. The next sections describe how systematic polysemous classes and underspecified semantic types can be derived from WordNet. In this paper I only consider classes of *nouns*, but the process described here can also be applied to other parts of speech.

2.1 Systematic polysemous classes

We can arrive at classes of systematically polysemous lexical items by investigating which items share the same senses and are thus polysemous in the same way. This comparison is done at the top levels of the WordNet hierarchy. WordNet does not have an explicit level structure, but for the purpose of this research one can distinguish a set of 32 'basic senses' that partly coincides with, but is not based directly on WordNet's list of 26 'top types':

> **act** (`act`), **agent** (`agt`), **animal** (`anm`), **artifact** (`art`), **attribute** (`atr`), **blunder** (`bln`), **cell** (`cel`), **chemical** (`chm`), **communication** (`com`), **event** (`evt`), **food** (`fod`), **form** (`frm`), **group_biological** (`grb`), **group** (`grp`), **group_social** (`grs`), **human** (`hum`), **linear_measure** (`lme`), **location** (`loc`), **location_geographical** (`log`), **measure** (`mea`), **natural_object** (`nat`), **phenomenon** (`phm`), **plant** (`plt`), **possession** (`pos`), **part** (`prt`), **psychological** (`psy`), **quantity_definite** (`qud`), **quantity_indefinite** (`qui`), **relation** (`rel`), **space** (`spc`), **state** (`sta`), **time** (`tme`)

Figure 3 shows their distribution among noun stems in the BROWN corpus. For instance there are 2550 different noun stems (with 49,824 instances) that have each 2 out of the 32 'basic senses' assigned to them in 238 different combinations (a subset of 32^2 = 1024 possible combinations).

senses	comb's	stems	instances
2	238	2550	49824
3	379	936	35608
4	268	347	22543
5	148	154	15345
6	52	52	5915
7	27	27	5073
8	10	10	3273
9	3	3	1450
10	1	1	483
11	2	2	959
12	1	1	441
	1161	10797	140914

Figure 3: Polysemy of nouns in BROWN

We now reduce all of WordNet's sense assignments to these basic senses. For instance, the seven different senses that WordNet assigns to the lexical item book (see Figure 1 above) can be reduced to the two basic senses: `'art com'`. We do this for each lexical item and then group them into classes according to their assignments.

From these one can filter out those classes that have only one member because they obviously do not represent a systematically polysemous class. The lexical items in those classes have a highly idiosyncratic behavior and are most likely homonyms. This leaves a set of 442 polysemous classes, of which Figure 4 gives a selection:

`act art evt rel`	click modification reverse
`act art log`	berth habitation mooring
`act evt nat`	ascent climb
`chm sta`	grease ptomaine
`com prt`	appendix brickbat index
`frm sta`	solid vacancy void
`lme qud`	em fathom fthm inch mil
`loc psy`	bourn bourne demarcation fairyland rubicon trend vertex
`log pos sta`	barony province
`phm pos`	accretion usance wastage
`rel sta`	baronetcy connectedness context efficiency inclusion liquid relationship

Figure 4: A selection of polysemous classes

Not all of the 442 classes are systematically polysemous. Consider for example the following classes:

`act anm art`	drill ruff solitaire stud
`act log`	bolivia caliphate charleston chicago clearing emirate michigan prefecture repair santiago wheeling
`act plt`	chess grapevine rape
`art fod loc`	pike port
`chm psy`	complex incense
`fod hum plt`	mandarin sage swede

Figure 5: A selection of *ambiguous* classes

Some of these classes are collections of homonyms that are *ambiguous* in similar ways, but do not lead to any kind of predictable polysemous behavior, for instance the class `'act anm art'` with the lexical items: drill ruff solitaire stud. Other classes consist of both homonyms and systematically polysemous lexical items like the class `act log`, which includes caliphate, clearing, emirate, prefecture, repair, wheeling vs. bolivia, charleston, chicago, michigan. Whereas the first group of nouns express two separated but related meanings (the **act** of clearing, repair, etc. takes place at a certain **location**), the second group expresses two meanings that are not related (the charleston dance which was named after the town by the same name).

The *ambiguous* classes need to be removed altogether, while the ones with mixed *ambiguous* and *polysemous* lexical items are to be weeded out carefully.

2.2 Underspecified semantic types

The next step in the research is to organize the remaining classes into knowledge representations that relate their senses to each other. These representations are based on Generative Lexicon theory ($\mathcal{GL}$), using *qualia roles* and *(dotted) types* (Pustejovsky, 1995).

Qualia roles distinguish different semantic aspects: FORMAL indicates semantic type; CONSTITUTIVE *part-whole* information; AGENTIVE and TELIC associated events (the first dealing with the *origin* of the object, the second with its *purpose*). Each role is typed to a specific class of lexical items. Types are either simple (**human, artifact,...**) or complex (e.g., **information•physical**). Complex types are called *dotted types* after the 'dots' that are used as type constructors. Here I introduce two kinds of dots:

> Closed dots '•' connect systematically related types that are always interpreted simultaneously.
>
> Open dots '∘' connect systematically related types that are not (normally) interpreted simultaneously.

Both '$\sigma \bullet \tau$' and '$\sigma \circ \tau$' denote sets of *pairs* of objects $\langle a, b \rangle$, a an object of type σ and b an object of type τ. A condition $a\mathbf{R}b$ restricts this set of pairs to only those for which some relation **R** holds, where **R** denotes a subset of the Cartesian product of the sets of type σ objects and type τ objects.

The difference between types '$\sigma \bullet \tau$' and '$\sigma \circ \tau$' is in the nature of the objects they denote. The type '$\sigma \bullet \tau$' denotes sets of pairs of objects where each pair behaves as a *complex* object in discourse structure. For instance, the pairs of objects that are introduced by the type **information•physical** (book, journal, scoreboard, ...) are addressed as the complex objects ⟨**x:information**, **y:physical**⟩ in discourse. On the other hand, the type '$\sigma \circ \tau$' denotes simply a set of pairs of objects that do not occur together in discourse structure. For instance, the pairs of objects that are introduced by the type **form•artifact** (door, gate, window, ...) are not (normally) addressed simultaneously in discourse, rather one side of the object is picked out in a particular context. Nevertheless, the pair as a whole remains active during processing.

The resulting representations can be seen as underspecified lexical meanings and are therefore referred to as *underspecified semantic types*. CoreLex currently covers 104 underspecified semantic types. This section presents a number of examples, for a complete overview see the CoreLex webpage:

http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html

Closed Dots Consider the underspecified representation for the semantic type **act•relation**:

$$\left[\begin{array}{l} \text{FORMAL} = \mathbf{Q{:}act{\bullet}relation} \\ \text{CONSTITUTIVE} = \\ \quad \mathbf{X{:}act} \vee \mathbf{Y{:}relation} \vee \mathbf{Z{:}act{\bullet}relation} \\ \text{TELIC} = \\ \quad \mathbf{P{:}event(act{\bullet}relation)} \wedge \mathbf{act(R_1)} \wedge \\ \quad \mathbf{relation(R_2,R_3)} \end{array}\right]$$

Figure 6: Representation for type: **act•relation**

The representation introduces a number of objects that are of a certain type. The FORMAL role introduces an object Q of type **act•relation**. The CONSTITUTIVE introduces objects that are in a part-whole relationship with Q. These are either of the same type **act•relation** or of the simple types **act** or **relation**. The TELIC expresses the **event** P that can be associated with an object of type **act•relation**. For instance, the event of increase as in 'increasing the communication between member states' implies 'increasing' both the **act** of communicating an object R_1 and the communication **relation** between two objects R_2 and R_3. All these objects are introduced on the semantic level and correspond to a number of objects that will be realized in syntax. However, not all semantic objects will be realized in syntax. (See Section 3.4 for more on the syntax-semantics interface.)

The instances for the type **act•relation** are given in Figure 7, covering three different systematic polysemous classes. We could have chosen to include only the instances of the `'act rel'` class, but the nouns in the other two classes seem similar enough to describe all of them with the same type.

`act evt rel`	blend competition flux transformation
`act rel`	acceleration communication dealings designation discourse gait glide likening negation neologism neology prevention qualifying sharing synchronisation synchronization synchronizing
`act rel sta`	coordination gradation involvement

Figure 7: Instances for the type: **act•relation**

Open Dots The type **act•relation** describes interpretations that can not be separated from each other (the **act** and **relation** aspects are intimately connected). The following representation for type **animal∘food** describes interpretations that can not occur simultaneously but are however related[1]. It therefore uses a '$\circ$' instead of a '$\bullet$' as a type constructor:

$$\begin{bmatrix} \text{FORMAL} = \textbf{Q:animal}\circ\textbf{food} \\ \text{CONSTITUTIVE} = \textbf{X:animal} \vee \textbf{Y:food} \\ \text{TELIC} = \\ \quad \textbf{P}_1\textbf{:act(R}_1\textbf{,animal)} \vee \textbf{P}_2\textbf{:act(animal,R}_2\textbf{)} \\ \quad \vee\ \textbf{P}_3\textbf{:act(R}_3\textbf{,food)} \end{bmatrix}$$

Figure 8: Representation for type: **animal∘food**

The instances for this type only cover the class `'anm fod'`. A case could be made for including also every instance of the class `'anm'` because in principal every animal could be eaten. This is a question of how *generative* the lexicon should be and if one allows *overgeneration* of semantic objects.

anm fod	bluepoint capon clam cockle crawdad crawfish crayfish duckling fowl grub hen lamb langouste limpet lobster monkfish mussel octopus panfish partridge pheasant pigeon poultry prawn pullet quail saki scallop scollop shellfish shrimp snail squid whelk whitebait whitefish winkle

Figure 9: Instances for the type: **animal∘food**

2.3 Homonyms

CoreLex is designed around the idea of systematic polysemous classes that exclude homonyms. Traditionally a lot of research in lexical semantics has been occupied with the problem of ambiguity in homonyms. Our research shows however that homonyms only make up a fraction of the whole of the lexicon of a language. Out of the 37,793 noun stems that were derived from WordNet 1637 are to be viewed as true homonyms because they have two or more *unrelated* senses, less than 5%. The remaining 95% are nouns that do have (an indefinite number of) different interpretations, but all of these are somehow related and should be inferred from a common knowledge representation. These numbers suggest a stronger emphasis in research on systematic polysemy and less on homonyms, an approach that is advocated here (see also (Killgariff, 1992)).

In CoreLex homonyms are simply assigned two or more underspecified semantic types, that need to be disambiguated in a traditional way. There is however an added value also here because each disambiguated type can generate any number of context dependent interpretations.

[1]See the literature on *animal grinding*, for instance (Copestake and Briscoe, 1992)

3 Adapting CoreLex to Domain Specific Corpora

The underspecified semantic type that CoreLex assigns to a noun provides a basic lexical semantic structure that can be seen as the class-wide backbone semantic description on top of which specific information for each lexical item is to be defined.

That is, doors and gates are both artifacts but they have different appearances. Gates are typically open constructions, whereas doors tend to be solid. This kind of information however is corpus specific and therefore needs to be adapted specifically to and on the basis of that particular corpus of texts.

This process involves a number of consecutive steps that includes the probabilistic classification of unknown lexical items:

1. Assignment of underspecified semantic tags to those nouns that are in CoreLex
2. Running class-sensitive patterns over the (partly) tagged corpus
3. (a) Constructing a probabilistic classifier from the data obtained in step 2.
   (b) Probabilistically tag nouns that are not in CoreLex according to this classifier
4. Relating the data obtained in step 2. to one or more qualia roles

Step 1. is trivial, but steps 2. through 4. form a complex process of constructing a corpus specific semantic lexicon that is to be used in additional processing for knowledge intensive reasoning steps (i.e. *abduction* (Hobbs et al., 1993)) that would solve metaphoric, metonymic and other non-literal use of language.

3.1 Assignment of CoreLex Tags

The first step in analyzing a new corpus involves tagging each noun that is in CoreLex with an underspecified semantic tag. This tag represents the following information: a definition of the type of the noun (FORMAL); a definition of types of possible nouns it can stand in a part-whole relationship with (CONSTITUTIVE); a definition of types of possible verbs it can occur with and their argument structures (AGENTIVE / TELIC). CoreLex is implemented as a database of associative arrays, which allows a fast lookup of this information in pattern matching.

3.2 Class-Sensitive Pattern Matching

The pattern matcher runs over corpora that are: part-of-speech tagged using a widely used tagger (Brill, 1992); stemmed by using an experimental system that extends the Porter stemmer, a stemming algorithm widely used in information retrieval, with the Celex database on English morphology; (partly) semantically tagged using the CoreLex set of underspecified semantic tags as discussed in the previous section.

There are about 30 different patterns that are arranged around the **headnoun** of an NP. They cover the following syntactic constructions that roughly correspond to a VP, an S, an NP and an NP followed by a PP:

- **verb-headnoun**
- **headnoun-verb**
- **adjective-headnoun**
- **modifiernoun-headnoun**
- **headnoun-preposition-headnoun**

The patterns assume NP's of the following generic structure[2]:

```
PreDet* Det* Num* (Adj|Name|Noun)* Noun
```

The heuristics for finding the **headnoun** is then simply to take the rightmost noun in the NP, which for English is mostly correct.

The **verb-headnoun** patterns approach that of a true 'verb-obj' analysis by including a normalization of passive constructions as follows:

```
[Noun Have?  Be Adv?  Verb] ⇒ [Verb Noun]
```

Similarly, the **headnoun-verb** patterns approach a true 'subj-verb' analysis. However, because no deep syntactic analysis is performed, the patterns can only approximate subjects and objects in this way and I therefore do not refer to these patterns as '**subject-verb**' and '**verb-object**' respectively.

The pattern matching is class-sensitive in employing the assigned CoreLex tag to determine if the application of this pattern is appropriate. For instance, one of the **headnoun-preposition-headnoun** patterns is the following, that is used to detect *part-whole* (CONSTITUTIVE) relations:

```
PreDet* Det* Num* (Adj|Name|Noun)* Noun of
PreDet* Det* Num* (Adj|Name|Noun)* Noun
```

Clearly not every syntactic construction that fits this pattern is to be interpreted as the expression of a part-whole relation. One of the heuristics we therefore use is that the pattern may only apply if both head nouns carry the same CoreLex tag or if the tag of the second head noun subsumes the tag of the first one through a dotted type. That is, if the second head noun is of a dotted type and the first is of one of its composing types. For instance, 'paragraph' and 'journal' can be in a *part-whole* relation to each other because the first is of type **information**, while the second is of type **information•physical**. Similar heuristics can be identified for the application of other patterns.

Recall of the patterns (percentage of nouns that are covered) is on average, among different corpora (WSJ, BROWN, PDGF – a corpus we constructed for independent purposes from 1000 medical abstracts in the MEDLINE database on Platelet Derived Growth Factor – and DARWIN – the complete *Origin of Species*), about 70% to 80%. Precision is much harder to measure, but depends both on the accuracy of the output of the part-of-speech tagger and on the accuracy of class-sensitive heuristics.

3.3 Probabilistic Classification

The knowledge about the linguistic context of nouns in the corpus that is collected by the pattern matcher is now used to classify unknown nouns. This involves a similarity measure between the linguistic contexts of classes of nouns that are in CoreLex and the linguistic context of unknown nouns. For this purpose the pattern matcher keeps two separate arrays, one that collects knowledge only on CoreLex nouns and the other collecting knowledge on *all* nouns.

The classifier uses *mutual information* (MI) scores rather than the raw frequences of the occurring patterns (Church and Hanks, 1990). Computing MI scores is by now a standard procedure for measuring the co-occurrence between objects relative to their overall occurrence. MI is defined in general as follows:

$$I\ (x\ y)\ =\ log_2\ \frac{P(x\ y)}{P(x)\ P(y)}$$

We can use this definition to derive an estimate of the *connectedness* between words, in terms of collocations (Smadja, 1993), but also in terms of phrases and grammatical relations (Hindle, 1990). For instance the co-occurrence of verbs and the heads of their NP objects (N: size of the corpus, i.e. the number of stems):

$$C_{obj}\ (v\ n)\ =\ log_2\ \frac{\frac{f(v\ n)}{N}}{\frac{f(v)}{N}\ \frac{f(n)}{N}}$$

All nouns are now classified by running a similarity measure over their MI scores and the MI scores of each CoreLex class. For this we use the *Jaccard measure* that compares objects relative to

[2] The interpretation of '*' and '?' in this section follows that of common usage in regular expressions: '*' indicates 0 or more occurrences; '?' indicates 0 or 1 occurrence

the attributes they share (Grefenstette, 1994). In our case the 'attributes' are the different linguistic constructions a noun occurs in: **headnoun-verb**, **adjective-headnoun**, **modifiernoun-headnoun**, etc.

The Jaccard measure is defined as the number of attributes shared by two objects divided by the total number of unique attributes shared by both objects:

$$\frac{A}{A + B + C}$$

A : attributes shared by both objects
B : attributes unique to object 1
C : attributes unique to object 2

The Jaccard scores for each CoreLex class are sorted and the class with the highest score is assigned to the noun. If the highest score is equal to 0, no class is assigned.

The classification process is evaluated in terms of precision and recall figures, but not directly on the classified unknown nouns, because their precision is hard to measure. Rather we compute precision and recall on the classification of those nouns that are in CoreLex, because we can check their class automatically. The assumption then is that the precision and recall figures for the classification of nouns that are known correspond to those that are unknown. An additional measure of the effectiveness of the classifier is measuring the recall on classification of *all* nouns, known and unknown. This number seems to correlate with the size of the corpus, in larger corpora more nouns are being classified, but not necessarily more correctly. Correct classification rather seems to depend on the homogeneity of the corpus: if it is written in one style, with one theme and so on.

Recall of the classifier (percentage of all nouns that are classified > 0) is on average, among different larger corpora ($>$ 100,000 tokens), about 80% to 90%. Recall on the nouns in CoreLex is between 35% and 55%, while precision is between 20% and 40%. The last number is much better on smaller corpora (70% on average). More detailed information about the performance of the classifier, matcher and acquisition tool (see below) can be obtained from (Buitelaar, forthcoming).

3.4 Lexical Knowledge Acquisition

The final step in the process of adapting CoreLex to a specific domain involves the 'translation' of observed syntactic patterns into corresponding semantic ones and generating a semantic lexicon representing that information.

There are basically three kinds of semantic patterns that are utilized in a CoreLex lexicon: hyponymy (sub-supertype information) in the FORMAL role, meronymy (part-whole information) in the CONSTITUTIVE role and predicate-argument structure in the TELIC and AGENTIVE roles. There are no compelling reasons to exclude other kinds of information, but for now we base our basic design on $\mathcal{GL}$, which only includes these three in its definition of qualia structure.

Hyponymic information is acquired through the classification process discussed in Sections 2.2 and 3.3. Meronymic information is obtained through a translation of various VP and PP patterns into 'has-part' and 'part-of' relations. Predicate-argument structure finally, is derived from **verb-headnoun** and **headnoun-verb** constructions.

The semantic lexicon that is generated in such a way comes in two formats: $\mathcal{TDL}$, a *Type Description Language* based on typed feature-logic (Krieger and Schaefer, 1994a) (Krieger and Schaefer, 1994b) and HTML, the markup language for the World Wide Web. The first provides a constraint-based formalism that allows CoreLex lexicons to be used straightforwardly in constraint-based grammars. The second format is used to present a generated semantic lexicon as a semantic index on a World Wide Web document. We will not elaborate on this further because the subject of semantic indexing is out of the scope of this paper, but we refer to (Pustejovsky et al., 1997).

3.5 An Example: The PDGF Lexicon

The semantic lexicon we generated for the PDGF corpus covers 1830 noun stems, spread over 81 CoreLex types. For instance, the noun evidence is of type **communication•psychological** and the following representation is generated:

$$\begin{bmatrix} \begin{array}{l} \textsf{evidence} \\ \\ \text{FORMAL} = \\ \quad \begin{bmatrix} \text{CLOSED} = \begin{bmatrix} \text{ARG1} = \textbf{communication} \\ \text{ARG2} = \textbf{psychological} \end{bmatrix} \end{bmatrix} \\ \text{CONSTITUTIVE} = \\ \quad \begin{bmatrix} \text{HAS-PART} = \begin{bmatrix} \text{FIRST} = \textbf{structure} \\ \text{REST} = \ldots \end{bmatrix} \end{bmatrix} \\ \text{TELIC} = \\ \quad \begin{bmatrix} \begin{array}{l} \text{FIRST} = \begin{bmatrix} \textbf{provide} \\ \text{ARG-STRUCT} = \ldots \end{bmatrix} \\ \\ \text{REST} = \ldots \end{array} \end{bmatrix} \end{array} \end{bmatrix}$$

Figure 10: Lexical entry for: evidence

4 Conclusion

In this paper I discuss the construction of a new type of semantic lexicon that supports *underspecified* semantic tagging. Traditional semantic tagging assumes a number of distinct senses for each lexical item between which the system should choose. Underspecified semantic tagging however assumes no finite lists of senses, but instead tags each lexical item with a comprehensive knowledge representation from which a specific *interpretation* can be *constructed*. CoreLex provides such knowledge representations, and as such it is fundamentally different from existing semantic lexicons like WordNet. Additionally, it was shown that CoreLex provides for more consistent assignments of lexical semantic structure among classes of lexical items. Finally, the approach described above allows one to generate domain specific semantic lexicons by enhancing Corelex lexical entries with corpus based information.

References

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL.

Buitelaar, Paul. forthcoming. CoreLex*: An Adaptable Semantic Lexicon with Systematic Polysemous Classes*. Ph.D. thesis, Brandeis University, Department of Computer Science.

Church, K. W. and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

Copestake, A. and E. Briscoe. 1992. Lexical operations in a unification-based framework. In James Pustejovsky and Sabine Bergler, editors, *Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence 627*, pages 22–29, Berlin. Springer Verlag.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 268–275.

Hobbs, J., M. Stickel, P. Martin, and D. Edwards. 1993. Interpretation as abduction. *Artificial Intelligence*, 63.

Killgariff, Adam. 1992. *Polysemy*. Ph.D. thesis, University of Sussex, Brighton.

Krieger, Hans-Ulrich and Ulrich Schaefer. 1994a. Tdl-a type description language for hpsg. part1: Overview. Technical Report RR-94-37, DFKI, Saarbruecken, Germany.

Krieger, Hans-Ulrich and Ulrich Schaefer. 1994b. Tdl-a type description language for hpsg. part2: Reference manual. Technical Report D-94-14, DFKI, Saarbruecken, Germany.

Pustejovsky, J., B. Boguraev, M. Verhagen, P.P. Buitelaar, and M. Johnston. 1997. Semantic indexing and typed hyperlinking. In *AAAI Spring 1997 Workshop on Natural Language Processing for the World Wibe Web*.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Pustejovsky, James, Bran Boguraev, and Michael Johnston. 1995. A core lexical engine: The contextual determination of word sense. Technical report, Department of Computer Science, Brandeis University.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1).

Stevenson, Mark and Yorick Wilks. 1997. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? To appear in a Special Issue of Computational Linguistics on Word sense Disambiguation.

Yarowsky, David. 1992. Word sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING92*, pages 454–460.