The Organization of the Lexicon: Semantic Types and Lexical Sets

Patrick Hanks
Department of Computer Science MS018,
Volen Center for Complex Systems,
Brandeis University,
Waltham MA 02454, USA

Abstract
This paper reports a new kind of lexicon currently being developed as a resource for natural language processing, language teaching, and other applications. This is a "Pattern Dictionary of English", based on detailed and extensive corpus analysis of each sense of each verb in the language. A pattern consists of a verb with its valencies, plus semantic values for each valency and other relevant clues, and is associated with an implicature that associates the meaning with the context rather than with the word in isolation. For each verb, all normal patterns are recorded. The semantic types in each argument slot are linked to actual words via a large ontology.
The paper discusses the relationship between A) words as they are actually used and B) semantic types and functions in a theoretical lexicon. An attempt will be made in the full paper to relate empirically observable, corpus-based facts about ordinary word use to the theoretical abstractions of Generative Lexicon Theory of James Pustejovsky and the Meaning-Text Theory of Igor Mel'èuk. (In an extended abstract, this can only be hinted at; the full paper will discuss it more fully.) Lexicography and linguistic theory are often uneasy bedfellows, but I shall suggest that in at least these two cases, there is a possibility of a harmonious and productive relationship.

1 Introduction

How is the lexicon of a language to be represented? The answer depends, of course, partly on the needs of a target user group, but also on the nature of the data. In this paper I focus on the needs of computational linguists, who are not well served by the lexical tools currently available to them. The work also has implications for language teaching and learning.

The lexicon of a language consists of a vast network of interrelated items, some more closely bound than others. Comparatively little is known about these interrelationships; what little there is has often been distorted in traditional presentations – thesauruses, dictionaries, and grammars – partly because of lack of evidence (up until about ten years ago) for how words actually behave and partly because of attempts to impose fashionable but inaccurate syntactically driven theoretical models on the lexicon. A satisfactory theory of a language (and of language in general) must be based on empirical analysis of what words do, not merely on the tail-end of top-down syntactic abstractions.

2 Ontologies

A thesaurus or "onomasiological dictionary" (à la Roget or WordNet) represents the content words of a language as a vast hierarchical ontology, with synonym sets that gloss over subtle distinctions. Such hierarchies are partly unexceptionable (who would deny that a ***canary*** IS-A ***bird***, or that a ***bird*** IS-A ***living being***?) and partly arbitrary or fictional (what grounds are there for believing that an ***idea*** **IS-A** ***concept***, rather than that, say a ***concept IS-A(n) idea***?). The commonsense organization of lexical items denoting physical objects appears to have been overextended in such ontologies and applied to all words, regardless of applicability, in disregard of empirical evidence. In particular, many abstract nouns do not fit comfortably into a hierarchical ontology. ***Information, description, explanation,*** **and** ***evidence***, for example, are abstract nouns in roughly the same semantic area, and they tend to occur in similar syntactic slots in relation to verbs such as ***give*** (that is, they have similar syntagmatic behaviour and they activate the same meaning of the verb), but this does not mean that they must be fitted into a hierarchy.

They seem to have similar or equal semantic values, so a flat grouping may be more appropriate.

In other cases, some semantic property other than hyponymy unites a lexical set: e.g. (in the terminology of Pustejovsky (1995)) the relevant feature of a set like *[chair, stool, settee, sofa, bench]* is its **telic** [for sitting on] not its **formal** [[Physical Object]] or its **constitutive** [has legs, a seat, a back,...] In other words, the relevant question sometimes is "What is it for?" not "What is its semantic type?" or "What is it made of?"

In other cases, the relevant question is axiological: "Is it good or bad?" For example, the main difference between ***encourage*** and ***incite*** is that you ***incite*** people to do bad things, while ***encourage*** is more neutral. A good ontology will group words together according to such properties if and only if they are supported by evidence of actual usage.

Up to now, in English lexicography, the syntagmatic aspect of language – collocations – the tendency of words to occur together, both in syntactically governed patterns and in unstructured proximities – has been somewhat neglected. This is all the more unfortunate if, as I believe, meanings can only be effectively attached to words in context, not to words in isolation.

3 Dictionaries: Traditional and Combinatorial

In traditional dictionaries, lexical items are listed alphabetically, and several statements (usually numbered statements) about a word's meaning are listed at each entry. One might expect that such a dictionary would say more than a thesaurus about the syntagmatics of words, for example showing how one meaning of a word is distinguished from another by its context, but this is rarely the case. An honourable exception is the French *Dictionnaire explicatif et combinatoire* of Mel'čuk and others (1984-). No comparable work exists in English.

What little is said in traditional English dictionaries is usually cautious and conservative – often restricted merely to top-level syntactic relations, and even those are inaccurate.

Thus, American "collegiate" dictionaries do not even recognize that a verb may have up to three arguments. They say no more than "transitive" and "intransitive", with occasional

mentions of distinctive prepositional choice. Thus, the verb ***put*** is described in such dictionaries only as a transitive verb. This would imply that "I put the cup" is a well-formed sentence of English. The notion that there might be an obligatory adverbial of place (e.g. "*Where* did you put the cup?") is not represented.

1. PATTERN: [[Person]] put [[PhysObj]] [Adv[Location]]
2. EXAMPLE: I put the cup down/{on the table}.

The syntagmatics of this sense of ***put*** are expressed formally in 1 and exemplified in 2.

This raises several question. If lexical sets are united by common properties, e.g. the semantic type [[PhysObj]], as suggested in 1, are those properties transferable to the same semantic types in relation to other verbs? I.e., is the set of physical objects that you can ***put somewhere*** the same as the set of physical objects that you can ***give*** or ***take*** or ***throw***?

I do not have a ready answer; it seems that some central members of a lexical set or semantic type occur repeatedly in relation to many verbs, while others drop out and new ones come in when the verb is changed.

In the case of the sense of ***put*** discussed here, which always has three arguments, the lexical sets populating each argument are vast, but nevertheless they are united by common semantic values, namely: Subject [[Person]], Direct Object [[Physical Object]], and Adverbial [[Location]].

3. EXAMPLE: The horse bolted.
4. PATTERN: {horse | rabbit | [[Person]] | [[Animate]]} bolts.

3. EXAMPLE: I bolted the stable door.
4. PATTERN [[Person]] bolts {door | window | gate | ...}

In the case of the verb ***bolt***, it is a prototype rather than a semantic type that unites the lexical items in at least two of the argument positions. The lexical set or semantic type is built around the prototype and the set may be unique to just one verb. Some mechanism is therefore needed to express this phenomenon. Stereotypically, it is a ***horse*** or ***rabbit*** that ***bolts,*** while if a human ***bolts*** something, it is most probably a ***door, gate,*** or ***window.***

It is not sufficient to use the semantic type [[Animate]] in the first case, because birds and cockroaches are [[Animate]] but they don't bolt). In the second case, the semantic type art [[Artefact]] would be similarly underrestrictive: dinner plates and TV sets are [[Artefact]]s but you don't bolt them.

In the case of the verb ***file,*** which traditional dictionaries define in terms of cataloguing papers or putting them away in an orderly fashion, it is necessary to state that if someone files a lawsuit, then that someone is the plaintiff or their lawyer and that, far from putting the papers away in orderly fashion, the lawyer lodges the papers with a court as a way of starting a procedure, namely a lawsuit. Even the best traditional dictionaries, which mention lawsuits in connection with ***file,*** do not correlate the subject and object in the way that is required if the meaning of the combination is to be processed correctly.

There are 12 syntagmatic patterns for ***file***, which summarize all normal uses of the verb. Most but not all of them are grouped around a stereotypical direct object with the broad semantic type [[Document]]. They include:

1 [[Person = Plaintiff | Lawyer]] files [[Document = Lawsuit]]
2 [[Person = Lawyer]] files [[Document = Evidence]]
3 [Person1]] files [[Document = Complaint]] (against [[Person2]])
4 [[Person = Judge]] files [[Document = Decision]]
5 [[Person = Taxpayer]] files [[Document = Tax Return]]
6 [[Person = Inventor]] files [[Document = Patent Application]]
7 [Person = Pilot]] files [[Document = Flight Plan]]
8 [[Person = Reporter]] files [[Document = Story]]
9 [[Person = Clerk]] files [PLURAL[Document]]

and, finally, some quite different senses:

10. [PLURAL[Person]] file [NO OBJ] [Adv[Direction]]
11. [[Person]] file {notch] {in [[PhysObj]]
12. [[Person]] file {[POSDET] nails}

The semantics of the arguments of each noun in the first nine patterns is determined, with more or less probability, by the verb and its other arguments. Thus, semantic types must be associated, on the one hand with lexical sets grouped around prototypical members, and on the other hand with semantic types realized in particular semantic roles. In some cases the semantic type is the relevant property that unites a group of words in a particular argument slot; in other cases some other property, for example the telic, is relevant. It is a task for future lexicographers to tease out the details of these complex relationships.

The first step is to identify, by corpus analysis, all the patterns of normal use associated with each verb. The verb is the pivot of the cluase, and many nouns will fall into place in a semantic ontology once their relationship – their normal relationship – to verbs is known. This is the current goal of the Pattern Dictionary project described in Hanks and Pustejovsky (2005) and elsewhere. The next step will be to identify the semantics of the nouns in argument slots and establish computationally to what extent groups of nouns recur in relation to different verbs. At this point, it will be possible to decide (or confirm) how an ontology, a collection of lexical relations, should really be organized in a way that is consistent with evidence of usage.

References

Hanks, P., Pustejovsky, J. (2005), 'A Pattern Dictionary for Natural Language Processing', in *Revue Française de Linguistique Appliquée* X: 2.

Mel'èuk, I. et al. (1984), *Dictionnaire explicatif et combinatoire du français contemporain. Montréal.* Les Presses de l'Université de Montréal.

Pustejovsky, J. (1995), *The Generative Lexicon.* Cambridge, MA, MIT Press.