Towards an empirically well-founded semantic ontology for NLP

Patrick Hanks
NLP Department, Faculty of Informatics
68a Botanická
602 00 Brno, Czech Republic
hanks@fi.muni.cz

Karel Pala, Pavel Rychly
NLP Department, Faculty of Informatics
68a Botanická
602 00 Brno, Czech Republic
{pala, pary}@fi.muni.cz

Abstract

This paper examines some issues involved in building a corpus-based ontology for use in determining the meaning of words in text, in the context of creating a "pattern dictionary". How do words cluster in paradigmatic lexical sets in actual usage (as reflected in a large corpus), and can these clusters be mapped onto a semantically structured ontology? What semantic notions need to be distinguished for this purpose, and what are the appropriate theoretical foundations? What other elements are needed for the application of determining meaning in text?

1 Introduction

It is a truism that the meaning of a word is, at least in part, determined by the contexts in which it is used. But what counts as context, and to what extent can the elements of contexts be encoded in an ontology? Before the advent of corpus linguistics, a traditional answer was that the possible contexts of words are so numerous and various that nothing useful can be said. While this may be true as far as it goes, it overlooks an important distinction, namely the distinction between all possible contexts and all normal contexts. Corpus analysis shows that, if we ask about all the *normal* contexts of a given word, then useful, distinctive, and measurable answers can be given—"distinctive" in the sense that very often relevant collocations determine or help to determine probabilistically the meaning of words in text. Typically, the nouns that occur in argument slots around a verb combine to determine the meaning of the verb. This can be illustrated with an example from Hanks and Pustejovsky (2005). Normally—prototypically—if a person ***fires*** something, the thing fired is either a firearm or another person. In the first case, the sense of the verb is 'to discharge a projectile from a firearm' and in the second case the sense is "dismiss from employment". The relevant patterns here are:

1. [[Human]] fire [[Firearm]]

and

2. [[Human 1]] fire [[Human 2]]

respectively. Related to pattern 1 is a third pattern:

3. [[Human]] fire [[Projectile]] from [[Firearm]]

The meaning of the verb in patterns 1 and 3 is roughly identical, although the arguments are different. However, even though bullets and guns are both artefacts, it does not make sense to conflate 1 and 3 into a hypothetical pattern such as 4, because there is a useful distinction to be made between the bullet, a [[Projectile]], which moves after the firing event, and the gun, a [[Firearm]], which stays put.

4. *[[Human]] fire [[Artifact]]

In the past, linguists devoted much attention to worrying about counterexamples derived from intuition-based invented scenarios—for example, a person being fired from a cannon in a circus. Although such a scenario is perfectly possible, it is not normal and it does not fit into a pattern of English usage. Moreover, even in abnormal contexts, the syntagmatic structure and the semantic values of other arguments very often give reasons for preferring one interpretation rather than another. Thus, a person fired from a cannon in a circus acquires ad hoc the semantic type value [[Projectile]]. So in 5 the normal semantic value of the lexical item *Jane*, [[Human]], is temporarily backgrounded under the influence the combination of the verb ***fire*** plus the directional adverbial argument "from a cannon". In this context, Jane becomes a sort of honorary [[Projectile]].

5. Jane was fired from a cannon yesterday.

Without the relevant adverbial argument, the verb cannot have this meaning. 6 can only mean that Jane lost her job.

6. Jane was fired yesterday.

If the direct object of ***fire*** is a projectile, the normal argument structure requires a source and/or goal. Conversely, if a source or goal is present in a sentence, the verb ***fire*** does not have the "dismiss from employment" meaning. The adverbial "from employment" may look like a source, but in fact it is not. In the Brandeis Semantic Ontology, *employment* is a [[Relational Process]] not a [[Location]] or an [[Artifact]]. A source must be a location or an

artifact, not a relational process.

A complicating factor is that if the direct object is a canonical member of the type [[Projectile]], the source and goal may be implied rather than explicit, as in 7.

> 7. Clegg denied that he had fired a bullet recovered from Miss Reilly's body.

The implication of all this is that, in principle, the meaning of clauses should be computable by reference to an inventory of the clausal norms, expressed in terms of argument structures with semantic type values (and certain other clues) extrapolated from corpus evidence, as in 1-3 above, but in practice the picture is complicated by various factors, for example the fact that clausal norms are exploited in various ways, for one reason and another (e.g. ellipsis and metaphor). Therefore, it is necessary to have a very clear theoretical basis and practical procedure for matching unseen sentences onto norms and for distinguishing exploitations of norms from the norms themselves.

The present paper discusses one aspect of this problem, namely the nature of an inventory of noun lexical sets in relation to the *Pattern Dictionary of English Verbs and their Arguments* currently being compiled at the Masaryk University, Brno. At the time of writing (April 2007), 1220 clausal patterns for 330 verbs have been compiled[1]. Each pattern is linked by the analyst to an "implicature", which expresses the meaning of the pattern. It is envisioned that applications will compare occurrences of verbs in texts to the patterns in the *Pattern Dictionary* and select the best match in order to find the meaning or implicature(s).

Work on the *Pattern Dictionary* proceeds verb by verb, rather than frame by frame. Up till now, automation has been deliberately avoided. This is because premature automation would have involved accepting as a given precisely those assumptions about language that we believe it is necessary to inquire into, in order to understand how words are used to make meanings. Our task is, among other things, to explore levels of generalization, based on close analysis of corpora, which will yield maximally informative implicatures—not to accept levels of generalization from an existing ontology. In this way, we aim to create a fundamental resource that will be of use for a variety of applications in computational linguistics, language learning, and cognitive psychology. With its focus on individual verbs, their patterns, and their different implicatures, rather than on situations or frames involving a range of semantically related words, the *Pattern Dictionary* may be expected in due course to provide a complementary resource to FrameNet, focusing on and distinguishing the different actual usages and meanings of each verb.

It should be added that, as soon as 5% of the work is completed (forecast for July 2007), the project will move into a new phase, exploring the possibility of using the regularities that have been observed so far as a basis for semi-automatic (but still interactive) sense-and-pattern distinctions.

Underlying inspirations for this work include preference semantics (Wilks, 1975) and generative lexicon theory (Pustejovsky, 1995). Unlike previous work on lexical ambiguity, the *Pattern Dictionary* starts, not with a predetermined list of possible meanings for each word derived from a machine-readable dictionary or from WordNet, but with a list of syntagmatic patterns derived from corpus analysis. Senses (implicatures) are linked to patterns, not to words in isolation. This greatly reduces the amount of ambiguity or 'lexical entropy' in the language as a system. Most (though not all) patterns are mutually exclusive. However, it raises other problems. The biggest single obstacle to rapid progress on the *Pattern Dictionary* at present is the absence of an ontology that will serve as a satisfactory tool for grouping nouns into lexical sets in argument slots in relation to particular verbs. This is not necessarily a criticism of ontologies: the goal of grouping concepts into a relational database and the goal of grouping words according to their syntagmatic behaviour may well be incompatible. This itself, however, would be an interesting finding.

2 The Nature of Ontologies Now

Currently existing ontologies include WordNet and the Brandeis Semantic Ontology (BSO). These are basically hierarchies of concepts that link hyponyms to their superordinates. Rumshisky et al. (2006) draw attention to some of the shortcomings of existing lexical resources such as WordNet and present the additional features of BSO in some detail. In this paper we discuss some of the problems that arise when attempting to use *any* ontology—be it BSO or

1 Approximately 550 additional patterns (100 verbs), compiled during a pilot study at Brandeis in 2004/5, are in a text file waiting to be updated and filtered into the database. The overall target of the project is to compile entries for all normal patterns for all normal verbs of English. On the basis of figures from the *Concise Oxford Dictionary*, we estimate this to be approximately 7,500 verbs and—extrapolating from work done so far—something in the order of 30,000 patterns. For comparison, it may be noted that WordNet offers a list of over 11,000 verbs, but on examination, this list turns out to contain several thousand ghost verbs such as *acidulate, advect, agroup, agnize, bedhop*, and *catenulate*, and phrases such as b*arrage jam, blanket jam,* and *cause to be perceived*. Whatever these are, they are not normal verbs of English.

WordNet—as a tool for distinguishing different senses and patterns of verbs.

BSO makes a fundamental distinction between [[Event]] and [[Entity]], whereas in WordNet events are entities. In BSO, a [[State]] (state of affairs) is an [[Event]] rather than an [[Entity]]. In GL terms, both these ontologies express formal and constitutive qualia, but WordNet neglects the telic and agentive, and both neglect syntagmatic relations.

Let us start by looking at a pair of simple lexical items: the noun ***dog*** and the verb ***bark***. Word Net 3.0 gives 186 hyponyms and subhyponyms for ***dog*** in the sense "member of the genus Canis", as opposed to "frump", "cad", "sausage" (*hot dog*), and "pawl". These hyponyms mostly denote breeds of dog (e.g. ***spaniel, dachshund, beagle***), but also include designations such as ***puppy, cur***, and ***mongrel. Bitch*** is a hyponym of the superordinate ***canine***. Looking upward in the WordNet hierarchy, we see that the same node (***dog***) participates in two formal trees:

> ***canine < carnivore < placental mammal < mammal < vertebrate < chordate < animal < organism < living thing < whole < object < physical entity < entity.***

and

> ***domestic animal < animal ...***

This second tree joins the first tree at the node ***animal***. (WordNet's category of domestic animals include cats as well as dogs). Several points may be made on the basis of this example.

- The nodes in WordNet's hierarchies are not all normal words of English. For example, ***placental mammal*** is a term of interest only to those concerned about about the status of the duck-billed platypus (which is an egg-laying and therefore non-placental mammal). ***Chordate*** is a term of even more arcane application. Apparently there exist some primitive marine creatures which have spinal columns (chordae) but not backbones (vertebrae). For almost all practical and NLP purposes, these organisms can be ignored: chordates are vertrebrates and vertebrates are chordates. Distinguishing between them is merely an scientific-obsessional distraction.
- The nodes in BSO seem, quite deliberately, to have names that (in many cases) distinguish them from anything that could be mistaken for a term of everyday English. They have names such as [[Human Agent of Activity]]. This emphasizes the fact that BSO is a conceptual hierarchy, independent of the lexicon of any language..
- The basis for these ontologies is the classificatory system of Western science as it has been developed from Aristotle and the European Enlightenment onwards, not the behaviour of words in actual usage in any language. This is a most important distinction, as it may help to explain some of the problems that arise when attempts are made to use current ontologies for disambiguation.
- There is no mention of syntagmatic relations such as that between the noun ***dog*** and the verb ***bark***. WordNet does offer a sentence frame for verbs. However, this does not express the syntagmatic relations of lexical items. For example, at ***bark*** the sentence frame is "Something barks" (no sign of a selectional preference for ***dog*** or even ***animal***). Some EuroWordNet projects (in particular Italian WordNet, but not English or German WordNets) show a limited number of syntagmatic relations systematically, for example between the noun ***cane*** 'dog' and the verbs ***abbaiare*** 'bark' and ***braccare*** 'hunt'. This represents a return to the practice of Wilkins (1668), who, in the words of Eco (1995), was "groping towards the modern notion of hypertext" when he added to his ontology the information that dogs bark and wolves howl. It is, however, a departure from the theoretical basis of Princeton WordNet, for such relations can only be represented satisfactorily on the basis of corpus-based preferences rather than the intuition-based certainties which underlie WordNet.
- WordNet's "synsets" are sometimes equated with separate senses of a word, but there is no good theoretical foundation for this equation. BSO does not lend itself so readily to such an equation.
- BSO's lexical items generally do not match very well with clusters found in a corpus. As intimated above, we have been asking ourselves why not.

BSO is work in progress, but at the time of writing it has a more streamlined ontological view of ***dog*** than WordNet. It places the scientific fact that dogs are mammals in a separate part of the hierarchy from the social fact that they are (normally) pets. Lexical items in BSO are plugged into the hierarchy at the appropriate place, generally near the bottom, like terminal nodes in a generative grammar. The relevant parts of the ontology for which ***dog*** is a terminal "lexical item" are as follows:

Dog < Mammal < Animal < Animate Living Entity < Organic Entity < Physical Object < Natural < Entity < TopType

and

Pet < Artifactual < Entity < TopType

There is a formal mapping between [[Pet]] and [[Animal]].

A main aim of the *Pattern Dictionary* is to show how each use of a verb is associated with a particular sense. So in the case of the verb ***bark,*** it distinguishes 1) *an animal barking* from 2) *people barking orders*, 3) *people barking up the wrong tree* (an idiom), and 4) *people barking their shins on a hard object* (the latter being a case where the verb is a homograph—a completely different lexical item, of different origin, which happens to be spelled identically).

For purposes of assigning the right meaning to the verb ***bark*** when we find it in a text, we can either take pot luck with percentages2 or we can use the *Pattern Dictionary*, in which case we need a list of lexical items denoting animals that (canonically) bark. A good start on this can be made with WordNet's 186 hyponyms of ***dog*** or with or BSO's lexical items under the semantic type [[Dog]] and its children. However, this does not tell the whole story.

3 Barking Muntjacs: a Rare Norm

So far, so good. But now we come to a couple of apparently peripheral problems that are in fact symptomatic of a much deeper problem. Not only dogs but also foxes, seals, baboons, and muntjacs bark. To represent this fact, the subject of the verb ***bark*** in the *Pattern Dictionary*, consists of a mixture of a semantic type and a set of lexical items:

{[[Dog]] | fox | seal | baboon | muntjac}

This inelegant formulation is necessary because, although going up the hyponym tree and down another branch of it will indeed find foxes, it will also find wolves, which (in the words of one journalist cited in the British National Corpus) "bark about as often as they appear at Crufts [dog show]". Moreover, although further excursions up and down the hyponym tree will retrieve for us seals, baboons, and muntjacs, it will also retrieve cats, elephants, and camels, which, of course, do not bark. It also, of course, raises the question, where should such excursions up and down hyponym trees stop?

This discussion illustrates why ontologies such as WordNet and BSO cannot be used as a means of organizing the lexical items of verb argument structure.

We hasten to add that acts of barking by seals, baboons, and muntjacs are extremely rare, even in large corpora,. compared with acts of barking by dogs and humans. The *Pattern Dictionary* has two options in cases like this. It can either decide that barking by muntjacs is an exploitation of the norm *[[Dog]] bark* or it can say that it is a rare but literal alternation of the norm. We treat it as a rare but literal alternation because, if muntjacs give voice at all, what they do is (quite literally) bark. This is a rare event in corpus texts because muntjacs are rare, but it is not a rare thing for muntjacs to do. If a muntjac or a seal gives voice at all, it barks. This is quite different from the case of the human cannonball. It is a norm of muntjac behaviour to bark, but it is not a norm of human behaviour to be fired from a cannon.

There is another side to the problem. Although it is typically the case that dogs bark, some dogs ***yap***, while others (or the same dogs in other circumstances) ***yelp*** or ***whine***. The *Pattern Dictionary* deals with this aspect of natural language use, firstly by making separate entries for all these verbs, each of which has at least one pattern with [[Dog]] as its prototypical subject, and secondly by adding secondary implicatures—for example, at the entry for Pattern 1 of ***bark*** a secondary implicature states that "Such cries are characteristic of adult large dogs".

4 The Nature of Lexical Sets

So far, we have seen that not everything that barks is a canine, while not everything that is a canine barks. This is a characteristic problem of syntagmatic lexical analysis. Existing ontologies, for all their undoubted merits, do not deal satisfactorily with this syntagmatic problem.

In the *Pattern Dictionary*, a lexical set is a paradigmatic cluster of words that activate the same sense of a verb and that have something in common semantically. Deciding what counts as "the same" sense of a verb and therefore what constitutes a member of a relevant lexical set cannot be done by rote procedure. Typically, the decision requires art

2 In a random sample of 100 uses of the verb ***bark*** from the British National Corpus, 62% of uses denote a dog barking; 23% denote a person barking words, 5% are the idiom "barking up the wrong tree". Conszideralbe sophistication is therefore requried to better than chance in identifying a rare pattern as diostinct from an indeterminate use of a common pattern.

and judgement, in particular about the appropriate level of generalization, bearing in mind the likely needs of an (unknown) user group. For example, in analysing the verb ***abate***, it is very clear that one large groups of uses involve a storm (prototypically) abating, with the implicature that its force diminishes, while another large group of uses involves a riot or other form of social unrest abating, with the implicature that society returns to a calmer and more ordered state. It may or may not be useful to make this distinction: it would be equally possible to lump them all together (with other uses too) in a general pattern "[[Problem]] abate". There is no single or obvious right answer to such dilemmas. There are, however, plenty of obvious wrong answers.

Lexical sets vary greatly in size. A lexical set may be very small, or it may be vast. Small lexical sets are associated in particular with light verbs and idioms, where a set of only one or two nouns may activate a distinctive meaning of the verb. So if the noun ***shins*** is found as the direct object of the verb ***bark,*** then the meaning of the sentence is very unlikely to have anything to with dogs making a noise or people saying things loudly. Moreover, although it is perfectly plausible to imagine other body parts in this particular slot (*ankles? knees? elbows? forearms*?), they do not occur as normal phraseology: both evidence and intuition militate against them. The idiom is, of course, always open to exploitations, but as part of the norm, these other body-part words are not found.

Slightly larger lexical sets are found with certain light verbs: so, prototypically, a person *takes a photograph,* but the variations *take a photo, take some snaps,* and *take a picture are* also found.

Much larger lexical sets (but still not vast) are those such as [[Firearm]]. These can be listed extensionally as well as defined intensionally. But there may be tension between an intensional definition and an otherwise plausible lexical set. For example, the perfectly normal, idiomatic phrase ***to fire an arrow*** plays havoc with any attempt at intensional definition of the verb ***fire*** in this sense. In many ways, firing an arrow activates the same implicatures as firing a bullet: a projectile is discharged from the 'firing' artifact; it moves at high velocity towards a target; it is intended to hit (and possibly damage) the target, and so on. It seems churlish to object that the mechanism of firing an arrow does not involve gunpowder. Yet, creating a separate category of the verb ***fire*** for the firing of arrows seems equally pointless or churlish.

It makes perfectly good sense to cluster together all words that normally denote projectiles and all the words that normally denote firearms. This must be a plausible basis for predicting other senses of verbs not yet analysed. But here a salient characteristic of lexical sets must be noted. As the set moves from verb to verb, some items drop out, while others come in. Thus, there is a statistically significant association in general-language corpora between words that denote firearms and the verb ***carry***. People carry a rifle, they carry a gun, they carry a revolver. But (for obvious practical reason) this association with ***carry*** does not extend to terms denoting large firearms such as ***cannon***. There are firearms that you carry and firearms that you do not carry. On the other hand, a firearm that cannot be fired would be a contradiction in terms, so the association between [[Firearm]] and the verb ***fire*** comes close to the Aristotelian doctrine of essences.

In some cases, a lexical set that picks out a particular sense of a verb may cut across several semantic types. For example, ***calm*** as a transitive (causative) verb has three or four main senses, including calming a riot, calming the stockmarket, and calming a person. With regard to the third of these senses, the direct objects are found in one lexical set but spread across five semantic types:

him, her, me, everyone: [[Human]]

fear, anger, temper, rage: [[Negative Feeling]]

mind: [[Psychological Entity]]

nerves, heart: [[Body Part]] (but not ***toes, chest,*** or ***kidney***)

breathing, breath: [[Living Entity Relational Process]] (but not ***defecation*** or ***urination***)

Thus, the relevant lexical set (words denoting a human attitude or emotional state) that activates this particular implicature for ***calm*** consists of a cluster of words drawn from a range of different semantic types—but these semantic types (as found in ontologies) quite rightly also contain many words that do not normally occur as direct object of ***calm***. Thus, an ontology of semantic types at best offers only a candidate list of possible items, not a full list of items that realize particular syntagmatic roles. In normal usage, only certain attributes of a human can be calmed.

5 Semantic Types vs. Semantic Roles

In the Pattern *Dictionary*, a distinction is made between a semantic type and a semantic role, as follows. The semantic type is an intrinsic attribute of a noun, while a semantic role has the attribute thrust upon it by the context. Thus, for example, pattern 1 (slightly tidied up) of the verb ***rule*** is:

[[Human = Monarch | Politician]] rule

[[Location | Human Group]]

We do not want to invoke a type [[Monarch]] or [[Politician]] as a semantic type here, partly because it implies a possible explosion of additional types in the ontology, and partly because such a move would make recognition and mapping in sentences such as 8 more difficult.

8. Blaize ruled Grenada for 10 unmemorable years.

There is nothing in the intrinsic semantics of ***Blaize*** to say that it denotes a politician or a monarch. In fact, as far as most readers are concerned, ***Blaize*** is no more than a name, most likely denoting some human being or other. The role [[Monarch]] or [[Politician]] is assigned by context—specifically, here, use as the grammatical subject of the verb ***rule***.

6 Attributes and Parts

Lexical sets contain not only synonyms and co-hyponyms, but also attributes. For example, a doctor may ***treat*** a patient (*pattern:* [[Human 1 = Health Professional]] treat [[Human 2 = Patient]]), but may also treat the patient's arm, knee, liver, etc. ([[Body Part]]), or an [[Ailment]] or [[Injury]] affecting the patient as a whole or just a particular body part.

In the *Pattern Dictionary*, lexical sets of this type are treated as semantic alternation, since not many of them transfer easily from verb to verb. However, aginst this (the notion of transferability of lexical sets), it must be pointed out that, in the last analysis, the lexical set for each argument of each verb is probably unique. This may not be so bad as it sounds, however, because onec the lexical set for each verb has been established with a number of canonical seed members, it is possible to add more members semi-automatically from corpus data.

Similarly, not only does something or someone ***scratch*** a physical object such as a piece of furniture, but also the ***surface*** or the ***leg*** or the ***top*** of the object. These are, of course, the most relevant parts of the furniture as far as getting scratched is concerned. Therefore, the lexical items ***surface, leg,*** and ***top*** must be in the same lexical set as the semantic type [[Furniture]] in relation to scratch, but not necessarily in relation to other verbs that take [[Furniture]] as an argument.

A further complication in the case of the verb ***scratch*** is that the expression *scratch the surface,* used absolutely, typically with a broad negatve (e.g. *we have hardly scratched the surface*), normally refers to the more obvious aspects of a complex problem or a mental entity, rather than to a physical object. If you want to talk about scratching the surface of a desk or other piece of furniture, you must mention the piece of furniture explictly somewhere in the co-text, usually but not always in the same noun phrase, after the preposition *of*.

7 Named Entities and Lexical Sets

Lexical sets must of course include names, and the names must be related to a semantic type. If we want to know what sort of event took place when Rex barked, we must know whether Rex is a dog or an irascible sergeant major. Sometimes, the problem is partly—i.e. probabilistically—solved by other components of a pattern. For example, if the verb ***bark*** governs direct speech, then Rex is much more probably a human than a dog. On the other hand, collocation with verbs such as ***snarl*** and ***bite*** increases the probability that Rex is a dog.

Named entities include not only places and people, but also business enterprises, horses, dogs, and other domestic animals, ships and even motor vehicles. Each of these sets has its own distinctive characteristics and members, though there is much overlap. A pattern and a name can interact to determine the semantic value of the verb, or the verb may help to decide what sort of entity a name most probably denotes. For example, the verb ***urge*** has at least two patterns:

[[Human 1]] urge [[Human 1]] {to [V]}

and

[[Human]] urge [[Horse]] [Adv[Direction]]

The second pattern applies mainly to horse, so in the sentence “Peter urged Bess up the lane”, Bess is most probably a horse, whereas in the sentence “Peter urged Bess to do it”, Bess is more probably a person. Even though we don't know what “it” was, we know that ***urge*** typically takes [[Human 2]] as direct object if it is further complemented by a *to*-infinitive. This kind of interaction between normal patterns and names and other aspects of language depends, of course, on being able to assign a name to the appropriate lexical set or semantic type.

8 Conclusion

The *Pattern Dictionary of English Verbs and their Arguments* is being created as a resource to help people map meaning onto words in use, use English idiomatically, and get a better understanding of the English language as a cognitive and social system. Parallel projects in other languages (Czech, Italian) are planned.

A pattern, in the sense used here, consists of a valency structure, which, basically, is any of several variations on the theme SVOA (Subject – Verb – Object – Adverbial). Each argument or valency consists of one or more lexical sets. Lexical sets are of two kinds: those that consist of only a few words (or even only one word in the case of certain idioms) and those that are so large that they must be derived from an ontology.

The main innovation of the *Pattern Dictionary* is that it starts by identifying patterns of use as found in a large corpus rather than starting from words in isolation. Only after the patterns have been distinguished from one another are they assigned meanings. Of course, this is an interactive procedure—the patterns are semantically motivated —and there are occasional residual ambiguities— but the focus on patterns rather than meanings reduces the ambiguity or lexical entropy of the language system to manageable proportions.

Considerably more interaction is needed between the *Pattern Dictionary* and other areas of linguistic research, for example parsing, anaphora resolution, and ontology building. In this paper, we have taken a brief look at the relationship between the *Pattern Dictionary* and ontologies.

The *Pattern Dictionary* currently focuses on the analysis of verbs and makes no predictions about the apparatus that will be needed for the analysis of nouns and attributive adjectives. However, it should be noted that already (after only 4% of English verbs have been analysed) many nouns are beginning to fall into place in the language system, insofar as their roles in relation to particular verbs are correctly identified on the basis of corpus pattern analysis. Unfortunately, as we have seen, they do not fall into place in quite the way that can be predicted on the basis of existing ontologies, no doubt because those ontologies were compiled with purposes other than corpus analysis of syntagmatic patterns in mind.

The absence of an existing empirically well-founded ontology that groups nouns together into paradigm sets according to their syntagmatic behaviour (as opposed to their place in a conceptual hierarchy) is a handicap for the *Pattern Dictionary* at present, so we are working on building our own shallow ontology, as well as on automatic identification of the paradigmatic clusters that constitute smaller lexical sets. An ontology of what Jackendoff has called "the semantic parts of speech"—[[Event]], [[State]]. [[Entity]], [[Human]], [[Physical Object]], [[Location]], etc. —is obviously an essential component of this research, coupled with a procedure for assigning words and names to nodes in an ontology. It is not possible to list extensionally all the [[Human]]s that there are, ever have been, or ever will be! So a generative procedure for identifying members of these large sets is a necessary complement of other work on lexical sets. It is an open question how far lexical sets of the kind discussed in this paper can be usefully organized into an ontology, and how far it will be more useful to leave them in open clusters. Because language is anthropocentric, [[Human]] is by far the most frequent and most important semantic type (even though in the great scheme of the universe, humanity may be utterly insignificant). A few other semantic types also have virtually open-ended membership, but after only about 100 very common semantic types, lexical sets are so much smaller that it may be more sensible to define them extensionally—i.e. by listing their most common members—rather than intensionally, by defining their attributes or essences.

9 Acknowledgements

This work has been partly supported by the Academy of Sciences of the Czech Republic under project T100300419 and by the Ministry of Education of the Czech Republic within the Center for basic research. LC536. and in the National Research Programme II project 2C06009.

References

Umberto Eco. 1995. *The Search for the Perfect Language*. Oxford: Blackwell.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique appliquée,* 10(2): 63-82.

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

Anna Rumshisky, Patrick Hanks, Catherine Havasi, and James Pustejovsky. 2006. Constructing a Corpus-based Ontology using Model Bias. *FLAIRS 2006*, Melbourne Beach, Florida.

John Wilkins. 1668. *Essay towards a Real Character, and a Philosophical Language.* London: The Royal Society.

Yorick Wilks. 1975. A Preferential, Pattern-Seeking, Semantics for Natural Language Inference. *Artificial Intelligence* 6(1): 53-74.

Web Site

Corpus Pattern Analysis and the Pattern Dictionary:
http://nlp.fi.muni.cz/projects/cpa/